Explainable, Trustable and Emphatic Artificial Intelligence from Formal Argumentation Theory to Argumentation for Humans

Serena Villata

UNIVERSITÉ CÔTE D'AZUR

HABILITATION THESIS

Habilitation á Diriger des Recherches (HDR)

Major: Computer Science

Serena Villata

EXPLAINABLE, TRUSTABLE AND EMPHATIC ARTIFICIAL INTELLIGENCE
FROM FORMAL ARGUMENTATION THEORY TO ARGUMENTATION FOR HUMANS

Jury:

Fabien Gandon, Research Director, INRIA (France), President
Leila Amgoud, Research Director, (IRIT, Toulouse) - Rapporteur
Simon Parsons, Professor, (King's College London, UK) - Rapporteur
Bernardo Magnini, Research Director, (FBK Trento, Italia) - Rapporteur

July 4th 2018

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank the main authors of the publications that are summarized here for their kind agreement to let these works contribute to this thesis.

Chapter 1

Introduction

This document relates and synthesizes my research and research management experience since I joined the EDELWEISS team led by Olivier Corby in 2011 for a postdoctoral position at Inria Sophia Antipolis. Then, in 2012, I got a "Starting Research Position" at Inria Sophia Antipolis in the WIMMICS (Web Instrumented Man-Machine Interactions, Communities and Semantics)[1], a joint team between Inria, University of Nice Sophia Antipolis and CNRS, led by Fabien Gandon[2]. In October 2015, I got a Researcher position (Chargé de Recherche – CR1) at CNRS. WIMMICS is a sub-group of the SPARKS[3] team (Scalable and Pervasive softwARe and Knowledge Systems) in I3S which has been structured into three themes in 2015. My research activity mainly contributes to the FORUM theme (FORmalising and Reasoning with Users and Models). Throughout this 10-year period, I was involved in several research projects and my role has progressively evolved from junior researcher to scientific leader of research actions. I initiated several research projects on my own, and I supervised several PhD thesis. In the meantime, I was also involved in the scientific animation of my research community (e.g., member of the CA of the AFIA association, chairing of conferences and workshops, PCs, ...).

My research area is Artificial Intelligence, with a particular interest in Knowledge Representation and Reasoning (KRR). More precisely, the majority of my works are in the area of argumentation theory. Argumentation theory is considered as a reasoning model based on the construction and evaluation of arguments. Arguments are supposed to support, contradict or explain statements, and they are used to support decision making. My background is a PhD in Artificial Intelligence, and more specifically in Argumentation Theory and Multiagent Systems, supervised by Guido Boella at the University of Turin (Italy) and Leendert van der Torre at the University of Luxembourg. During my PhD, I worked on the definition of the methodology and techniques of meta-argumentation to model argumentation. The methodology of meta-argumentation instantiates Dung's abstract argumentation theory with an extended argumentation theory, and it is thus based on a combination of the methodology of instantiating abstract arguments, and the methodology of extending Dung's basic argumentation frameworks with other relations among abstract arguments. The technique of meta-argumentation applies Dung's theory of abstract argumentation to itself, by instantiating Dung's abstract arguments with meta-arguments using a technique called *flattening*. I characterized the domain of instantiation using a representation technique based on soundness and completeness. I applied this innovative technique to three different kinds of reasoning problems: (i) reasoning about support in bipolar

[1] In 2010, the Inria EDELWEISS team and the I3S KEWI team merged to become WIMMICS.

[2] `http://wimmics.inria.fr/`

[3] `http://sparks.i3s.unice.fr/`

abstract argumentation, (ii) reasoning about trust in multiagent systems, and (iii) reasoning about coalitions in multiagent systems.

After my PhD, my research topics gradually evolved from formal computational models of argument to natural models of arguments, where arguments in natural language are considered and mined from heterogeneous sources. I am one of the very first initiators of the research topic, very popular nowadays, called Argument Mining. I have also studied the impact of emotions on the argumentation online debaters express on the Web. These two research lines are highly multidisciplinary, and they have been addressed thanks to a fruitful collaboration with linguists and cognitive scientists. The rationale behind my research work (and more generally behind the whole research activity of the WIMMICS team) is that decision (being them from artificial or human agents) need to be supported and justified through arguments and other factors, like emotions, mental states and trust, and the Web represents a invaluable interaction architecture and information source to be leveraged in order to support decision making and the subsequent justification about such deliberation.

1.1 Research areas

In this section, I provide an overview of the three main research areas my contributions deal with, i.e., I contribute to the areas of computational models of argument, argument mining, and normative reasoning.

Computational Models of Argument

In everyday life, arguments are "reasons to believe and reasons to act". The idea of "argumentation" as the process of creating arguments for and against competing claims, has long been a subject of interest to philosophers and lawyers. In recent years, however, there has been a growth of interest in the subject from formal and technical perspectives in Computer Science, and a wide use of argumentation technologies in practical applications. In Computer Science, argumentation is viewed as a mechanical procedure for interpreting events, organizing and presenting documents and making decisions about actions. From a theoretical perspective, argumentation offers a novel framework casting new light on classical forms of reasoning, such as logical deduction, induction, abduction and plausible reasoning, communication explanations of advice. These forms of reasoning support discussion and negotiation in computer-supported cooperative work, and learning. From a human-computer interaction point of view, argumentation is a versatile technique that facilitates natural system behavior and is more easily understood by human users and operators. Argumentation theory involves different ways for analyzing arguments and their relationships. In particular, my research focuses on abstract argumentation proposed by Dung in 1995 that sees each argument as an abstract entity and in which arguments are related to each other by means of attack relations.

Complex technical systems and services increasingly require several autonomous agents that have to collaborate and communicate in order to achieve required objectives, because of the inherent interdependencies and constraints that exist between their goals and tasks. Increasingly they depend upon complex conversations concerned with negotiation, persuasion and trustworthiness where agents have different capabilities and viewpoints. Such dialogues have at their heart an exchange of proposals, claims or offers. What distinguishes argumentation-based discussions from other approaches is that proposals can be justified by the arguments that support, or oppose, them. This permits greater flexibility than in other decision-making and communication schemes since, for instance, it makes it possible to persuade agents to change their view of a claim by identifying information or knowledge that is not being considered, or by introducing a new relevant factor in the middle of a negotiation or to resolve an impasse.

Argumentation is the process by which arguments are constructed and handled. Thus argumentation means that arguments are compared, evaluated in some respect and judged in order to establish whether any of them are warranted. Each argument is a set of assumptions that, together with a conclusion, is obtained by a reasoning process [42]. The layout of an argument has been studied by Toulmin in 1958 [302] who identified the pieces of information composing an argument. These key components are the data, the claim, the warrant and the rebuttal. A claim is a conclusion which is drawn if the warrant holds and the rebuttal does not hold. The data, supported by the warrant, imply the claim. Argumentation as exchange of pieces of information and reasoning about them involves groups of agents. We can assume that each argument has at least a proponent, the person who puts forward the argument, and an audience, the person who receives the argument. Two kinds of views on argumentation can be highlighted in multiagent systems, monological and dialogical. In the former, a single agent or a group of agents with the same role has the knowledge to construct arguments to support and attack a conclusion while, in the latter, a group of agents interacts to construct arguments supporting or attacking a particular claim.

There are, at the higher level, two ways to formalize a set of arguments and their relationships, abstract argumentation and logical argumentation. Abstract argumentation has been introduced by Dung [133], and it names only the arguments without describing them at all; it just represents that an argument is attacked by another one. Logical argumentation [2, 42, 263] is a framework in which more details about the arguments are considered. In particular, each argument is seen as composed by the premises, the claim and the inference rules used to achieve the claim from the premises.

My past research activity concentrated mainly on Dung-like abstract argumentation systems. The motivation of the meta-argumentation methodology proposed in my PhD thesis comes from the well known and generally accepted observation that Dung's theory of abstract argumentation cannot be used directly when modeling argumentation in many realistic examples, such as multiagent argumentation and dialogues, decision making, coalition formation, combining Toulmin's micro arguments, normative reasoning. The development of the meta-argumentation methodology is the main contribution of my PhD thesis. My key idea was that meta-argumentation instantiates Dung's theory with meta-arguments, such that Dung's theory is used to reason about itself. In my thesis, I applied the methodology of meta-argumentation to three challenges that in recent years have involved the research area of argumentation theory: reasoning about support relations among arguments, reasoning about trust, and reasoning about coalitions of agents.

After the thesis, I exploited my results in representing bipolar argumentation using meta-argumentation in two different fields which offer challenging open issues to reason over the acceptability of the arguments: inconsistencies detection in requirements engineering[4] (together with Isabelle Mirbel (UNS), *Isabelle Mirbel, Serena Villata. Enhancing Goal-Based Requirements Consistency: An Argumentation-Based Approach. 13th International Workshop on Computational Logic in Multi-Agent Systems (CLIMA 2012): 110-127*), and in question-answering over the Web (together with Elena Cabrio (UNS), *Elena Cabrio, Serena Villata, Alessio Palmero Aprosio. A RADAR for information reconciliation in Question Answering systems over Linked Data. Semantic Web Journal 8(4): 601-617 (2017)*).

Concerning reasoning about trust, one challenge I tackled in my contributions was to use argumentation theory not only to model whether an information source is trusted or not, but also to understand the reasons, modeled under the form of arguments, for trusting the sources in case of conflicts concerning their trustability. In this line of work, I started from the meta-argumentation methodology I defined in my PhD thesis, and I extended it to be applied to reasoning about conflicts in trust. This contribution has been published at the ECSQARU conference (*Serena Villata, Guido Boella, Dov M. Gabbay, Leendert van der Torre.*

[4] `http://www-sop.inria.fr/members/Serena.Villata/argRE.html`

Arguing about the Trustworthiness of the Information Sources. 11th European Conference Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2011): 74-85) and on the International Journal on Approximate Reasoning (*Serena Villata, Guido Boella, Dov M. Gabbay, Leendert van der Torre. A socio-cognitive model of trust using argumentation theory. Int. J. Approx. Reasoning 54(4): 541-559 (2013)*).

In addition, I proposed (together with Celia da Costa Pereira (UNS) and Andrea Tettamanzi (UNS)) a fuzzy labeling algorithm to assign arguments in an abstract argumentation framework with an acceptability degree depending on the trustworthiness of the source proposing them. This contribution has been published at the IJCAI and SUM conferences: *Célia da Costa Pereira, Andrea Tettamanzi, Serena Villata. Changing One's Mind: Erase or Rewind? 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011): 164-171* and *Célia da Costa Pereira, Mauro Dragoni, Andrea Tettamanzi, Serena Villata. Fuzzy Labeling for Abstract Argumentation: An Empirical Evaluation. 10th International Conference on Scalable Uncertainty Management (SUM 2016): 126-139.* This line of work has been continued in the PhD thesis of Amel Ben Othmane [34, 33, 32], I supervised with Andrea Tettamanzi and Nhan Le Than. The title of the thesis is "CARS - A multi-agent framework to support the decision making in uncertain spatio-temporal real-world applications", and Amel successfully defended it on October 12th, 2017.

Finally, in the area of computational models of arguments, I also proposed a formal approach to modularity and decomposability of argumentation frameworks. I studied argumentation frameworks as interacting components, characterized by an Input/Output behavior, rather than as isolated monolithic entities. This modeling stance is particularly relevant in applications like argument summarization and explanation. Together with Massimiliano Giacomin and Pietro Baroni from the University of Brescia (Italy), we have started by introducing a general modeling approach and providing a comprehensive set of theoretical results putting the intuitive notion of Input/Output behavior of argumentation frameworks on a solid formal ground. This contribution has been published on the Artificial Intelligence Journal (*Pietro Baroni, Guido Boella, Federico Cerutti, Massimiliano Giacomin, Leendert van der Torre, Serena Villata. On the Input/Output behavior of argumentation frameworks. Artif. Intell. 217: 144-197 (2014)*).

In 2015, I also organized, together with Matthias Thimm (Koblenz University) the First International Competition on Computational Models of Argumentation (ICCMA 2015). In this competition, different solvers have been compared over the resolution of a set of abstract argumentation frameworks with respect to standard Dung semantics. The analysis of the results has been published on the Artificial Intelligence Journal (*Matthias Thimm, Serena Villata. The first international competition on computational models of argumentation: Results and analysis. Artif. Intell. 252: 267-294 (2017)*).

Together with many other colleagues in the area of Computational Models of Arguments, I have also authored a popularization paper for AI Magazine with an overview of the research area challenges (*Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, Serena Villata. Towards Artificial Argumentation. AI Magazine 38(3): 25-36 (2017)*).

Argument Mining

In the last years, the growing of the Web and the daily increasing number of textual data published there with different purposes have highlighted the need to process such data in order to identify, structure and summarize this huge amount of information. Online newspapers, blogs, online debate platforms and social networks, but also normative and technical documents provide an heterogeneous flow of information where natural language arguments can be identified, and analyzed. The availability of such data, together with

the advances in Natural Language Processing and Machine Learning, supported the rise of a new research area called *argument mining*. The main goal of argument mining is the automated extraction of natural language arguments and their relations from generic textual corpora, with the final goal to provide machine-readable structured data for computational models of argument and reasoning engines. Together with Elena Cabrio, I was one of the initiators of this research field, with a first paper on argument mining published at ECAI-2012 (*Elena Cabrio, Serena Villata. Natural Language Arguments: A Combined Approach. 20th European Conference on Artificial Intelligence (ECAI 2012): 205-210*) and the organization of the workshop "Frontiers and Connections between Argumentation Theory and Natural Language Processing" in July 2014.

Two main stages have to be considered in the typical argument mining pipeline, from the unstructured natural language documents towards structured (possibly machine-readable) data:

Arguments' extraction : The first stage of the pipeline is to detect arguments within the input natural language texts. Referring to standard *argument graphs*, the retrieved arguments will thus represent the nodes in the final argument graph returned by the system. This step may be further split in two different stages such as the extraction of arguments and the further detection of their boundaries. Many approaches have recently tackled such challenge adopting different methodologies like for instance Support Vector Machines [248, 250, 290, 141, 207], Naive Bayes classifiers [44], Logistic Regression [198].

Relations' extraction : The second stage of the pipeline consists in constructing the argument graph to be returned as output of the system. The goal is to predict what are the relations holding between the arguments identified in the first stage. This is an extremely complex task, as it involves high-level knowledge representation and reasoning issues. The relations between the arguments may be of heterogeneous nature, like attack, support or entailment [70]. This stage is also in charge of predicting, in structured argumentation, the internal relations of the argument's components, such as the connection between the premises and the claim [44, 290]. Being it an extremely challenging task, existing approaches assume simplifying hypotheses, like the fact that evidence is always associated with a claim [1]

To address these issues, solve problems and build applications upon, tools must be developed to analyze, aggregate, synthesize, structure, summarize, and reason about arguments in texts. However, to do so, more linguistic sophistication is required as well as newer techniques than currently found in Natural Language Processing.

Moreover, to tackle these challenging tasks, high-quality annotated corpora are needed, as those proposed in [271, 248, 198, 1, 290, 73, 164], to be used as a training set for any kind of aforementioned prediction. These corpora are mainly composed by three different elements: an *annotated* dataset which represents the gold standard whose annotation has been checked and validated by expert annotators and is used to train the system for the required task (i.e., arguments or relations extraction), a set of guidelines to explain in a detailed way how the data has been annotated, and finally, the *unlabelled* raw corpus that can be used to test the system after the training phase. The reliability of a corpus is assured by the calculation of the inter-annotator agreement that measures the degree of agreement in performing the annotation task among the involved annotators.[5] Current prototypes of argument mining systems require to be trained against the

[5]The number of involved annotators should be > 1 in order to allow for the calculation of this measure and, as a consequence, produce a reliable resource.

data the task is addressed to, and the construction of such annotated corpora remains among the most time consuming activities in this pipeline.

Together with Elena Cabrio (UNS), I proposed a combined framework of natural language processing and argumentation theory to support human users in their interactions. The framework combines a natural language processing module which exploits the Textual Entailment (TE) approach and detects the arguments in natural language debates and the relationships among them, and an argumentation module which represents the debates as graphs and detects the accepted arguments. The argumentation module is grounded on bipolar argumentation, and on the results previously achieved during my PhD thesis. Moreover, I studied the relation among the notion of support in bipolar argumentation, and the notion of TE in Natural Language Processing (NLP). The results of this research has been published both in the Computational Models of Argument venues (*Elena Cabrio, Serena Villata. Generating Abstract Arguments: A Natural Language Approach. Computational Models of Argument (COMMA 2012): 454-461* and *Elena Cabrio, Serena Villata. A natural language bipolar argumentation approach to support users in online debate interactions. Argument & Computation 4(3): 209-230 (2013)*) and in Natural Language Processing ones (*Elena Cabrio, Serena Villata. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012): 208-212*).

This research line about natural models of argumentation resulted in the NoDE Benchmark[6], a benchmark of natural arguments extracted from different kinds of textual sources. It is composed of three datasets of natural language arguments, released in two machine-readable formats, i.e., the standard XML format, and XML/RDF format adopting the SIOC-Argumentation vocabulary (extended). Arguments are connected by two kinds of relations: a positive (i.e., support) relation, and a negative (i.e., attack) relation, leading to the definition of bipolar argumentation graphs.

I started also to investigate the mapping between argumentation schemes in argumentation theory, and discourse in Natural Language Processing, together with Elena Cabrio (UNS) and Sara Tonelli (FBK Trento).

Since 2012, I continued to work on argument mining, facing the challenges raised by different domains, i.e., social media like Twitter (publications on this topic: *Tom Bosc, Elena Cabrio, Serena Villata. Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media. Computational Models of Argument (COMMA 2016): 21-32*, *Tom Bosc, Elena Cabrio, Serena Villata. DART: a Dataset of Arguments and their Relations on Twitter. Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, and *Mihai Dusmanu, Elena Cabrio, Serena Villata. Argument Mining on Twitter: Arguments, Facts and Sources. 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017): 2317-2322*), political speeches in collaboration with Sara Tonelli and Stefano Menini from FBK-Trento (publication on this topic: *Stefano Menini, Elena Cabrio, Sara Tonelli, Serena Villata. Never Retreat, Never Retract: Argumentation Analysis for Political Speeches. Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*), and legal cases in collaboration with Laura Alonso Alemany, Cristian Cardellino and Milagro Teruel from the University of Cordoba in Argentina (publication on this topic: *Milagro Teruel, Cristian Cardellino, Laura Alonso Alemany, Serena Villata. Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines. 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*).

[6] `http://www-sop.inria.fr/NoDE/`

Normative Reasoning

Normative systems are systems in the behavior of which norms play a role and which need normative concepts in order to be described or specified. A normative multi-agent system combines models for normative systems (dealing for example with obligations, permissions and prohibitions) with models for multi-agent systems. Normative multi-agent systems provide a promising model for human and artificial agent coordination because they integrate norms and individual intelligence. In particular, I adopt deontic logic to reason about licenses compatibility and composition. Deontic logic is the study of the logical relations among propositions that assert that certain actions or states of affairs are obligatory, forbidden or permitted.

The Web is evolving from an information space for sharing textual documents into a medium for publishing structured data too. The Linked Data initiative aims at fostering the publication and interlinking of data on the Web, giving birth to the so called Web of Data, an interconnected global dataspace where data providers publish their content publicly or under open licenses. Heath and Bizer underline that "the absence of clarity for data consumers about the terms under which they can reuse a particular dataset, and the absence of common guidelines for data licensing, are likely to hinder use and reuse of data". Therefore, all Linked Data on the Web should include explicit licensing terms. The explicit definition of the licensing terms under which the data is released is an open problem both for the data provider and for the data consumer. The former needs to explicit the licensing terms to ensure use and reuse of the data compliant with her requirements. The latter, instead, needs to know the licenses constraining the released data to avoid misusing and even illegal reuse of such data.

In the context of the Datalift project, I defined a fine-grained context-aware access control model for the Web of Data. I introduced the *Social Semantic SPARQL Security for Access Control* vocabulary `S4AC`[7], a lightweight ontology which allows to define fine-grained access control policies for RDF data.

Concerning licensing information, I proposed, together with Guido Governatori (Data61 - CSIRO) and Antonino Rotolo (University of Bologna), a formal framework based on deontic logic to verify the compatibility of a set of access rights and composed them into a unique set of access rights to be associated to a query result. This contribution has been published in AI & Law conferences: *Antonino Rotolo, Serena Villata, Fabien Gandon. A deontic logic semantics for licenses composition in the web of data. International Conference on Artificial Intelligence and Law (ICAIL 2013): 111-120* and *Guido Governatori, Ho-Pun Lam, Antonino Rotolo, Serena Villata, Fabien Gandon. Heuristics for Licenses Composition. Twenty-Sixth Annual Conference Legal Knowledge and Information Systems (JURIX 2013): 77-86*. Then, through a fruitful collaboration with the University of Cordoba, this framework has been associated to a new module, which has the aim to move from natural language formulations of the licenses to the machine readable ones (using RDF). This research line resulted in the design and development of the Licentia suite of services[8], to support users to choose the best license for their data. Still in collaboration with the University of Cordoba, we are currently working on a Named Entities Recognizer, Classifier and Linker for the legal domain so as to be able to support Information Extraction methods applied to legal documents. The results of this research line have been published in Natural Language Processing venues and AI & Law ones: *Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, Serena Villata. A low-cost, high-coverage legal named entity recognizer, classifier and linker. 16th edition of the International Conference on Artificial Intelligence and Law (ICAIL 2017): 9-18*, *Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, Serena Villata. Legal NERC with ontologies, Wikipedia and curriculum learning. 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017): 254-259*, and *Cristian Cardellino, Milagro Teruel,*

[7] `http://ns.inria.fr/s4ac/`

[8] `http://licentia.inria.fr/`

Laura Alonso Alemany, Serena Villata. Learning Slowly To Learn Better: Curriculum Learning for Legal Ontology Population. Thirtieth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2017): 252-257.

1.2 Application Domains and Research Projects

In this section, I report about my involvement in research projects in the above mentioned research areas.

Argumentation and Argument Mining

Regarding my activity in the argument mining domain applied to social media, I was the co-leader (together with Elena Cabrio (UNS)) for Inria of the CARNOT Project (2014-2015) with the start-up Vigiglobe. The goal of this project was to apply the argument mining pipeline to Twitter data. Understanding and interpreting the flow of messages exchanged in real time on social platforms, like Twitter, raises several important issues. The big amount of information exchanged on these platforms represents a significant value for who is able to read and enrich this multitude of information. Users directly provide this information, and it is interesting to analyze such data both from the quantitative and from the qualitative point of view, especially for what concerns reputation and marketing issues (regarding brands, institutions or public actors). Moreover, the automated treatment of this type of data and the constraints it presents (e.g., limited number of characters, tweets with a particular writing style, amount of data, real-time communication) offer a new and rich context for a challenging use of existing tools for argument mining. In the context of this project, I supervised the activity of Tom Bosc, a research engineer, now doing a PhD in Canada. Also in the context of this project, I supervised the 3-month internship of Mihai Dusmanu (École normale supérieure, Paris) about Argument Detection on Twitter. The publications related to this research line have been listed above.

My research activity dealing with emotions in argumentation started with the SEEMPAD[9] project (Joint Research Team with Heron Lab – 2014-2016). The goal of this project was to study the different dimensions of exchanges arising on online discussion platforms. More precisely, we concentrated on the study and analysis of the impact of emotions and mental states on the argumentation in online debates, with a particular attention to the application of argumentative persuasion strategies. The results of this research line have been published both in AI venues (*Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, Fabien Gandon. Emotions in Argumentation: an Empirical Evaluation. Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015): 156-163*), in Cognitive Science venues (*Sahbi Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, Elena Cabrio. Persuasive Argumentation and Emotions: An Empirical Evaluation with Users. 19th International Conference on Human-Computer Interaction (HCI 2017): 659-671*), and on the Argument & Computation journal (*Serena Villata, Elena Cabrio, Iméne Jraidi, Sahbi Benlamine, Maher Chaouachi, Claude Frasson, Fabien Gandon. Emotions and personality traits in argumentation: An empirical evaluation1. Argument & Computation 8(1): 61-87 (2017)*).

Again about argument mining, I am the French co-leader (together with Elena Cabrio) of the EIT CREEP (Cyberbulling Effect Prevention – 2018-2019) project. The purpose of CREEP is to provide a set of tools to support the detection and prevention of psychological/behavioral problems of cyberbullying teenage victims. The objective will be achieved combining social media monitoring and motivational technologies (virtual coaches integrating chatbots). Starting from February 2018, I supervise the activity of two research

[9] https://project.inria.fr/seempad/

engineers, Pinar Arslan and Michele Corazza, working on mining arguments where cyberbullism features are identified.

Moreover, I co-supervise with Leendert van der Torre the PhD thesis (2017 - 2020) of Shohreh Haddadan on argument mining in political debates.

I am also currently involved in the UCA IADB project (Intégration et Analyse de Donn'ees Biomédicales – 2017-2020). The goal of the project is to define the methods to access, process and extract information from a huge quantity of biomedical data from heterogeneous sources and of different nature (e.g., texts, images, ...). In the context of this project, I supervise (together with Johan Montagnat and Celine Poudat) the PhD thesis of Tobias Mayer about "Argument Mining on Clinical Trials", starting from October 2017.

I am also involved in the Programme d'Investissements d'Avenir "Grans Defis Du Numerique Big Data" ANSWER Project (Advanced aNd Secured Web Experience and seaRch – 2018-2020). The goal of the project is to build the next generation of search engines. My involvement is related to the supervision of the PhD thesis of Vorakit Vorakitphan about emotion-based argument detection, to go beyond standard sentiment analysis techniques, and empower search engines with the ability to deal with emotional criteria as well.

Normative reasoning

My post-doc at Inria Sophia Antipolis was in the context of the ANR Datalift project (2010-2013). The goal of DataLift was to provide a complete path from raw data to fully interlinked, identified, qualified and "certified" linked data sets. The resulting Datalift platform supported the processes of selecting ontologies for publishing data; converting data to the appropriate format (RDF using the selected ontology); interlinking data with other data sources; publishing linked data. My task in the project was the definition of a right expression and management module for the Web of Data.

In the context of normative reasoning, I am the French principal investigator of the European Union's Horizon 2020 research and innovation programme Marie Sklodowska-Curie MIREL project (Mining and Reasoning with Legal Texts – 2016-2019). The goal of MIREL is to create an international and intersectorial network to define a formal framework and to develop tools for mining and reasoning with legal texts, with the aim of translating these legal texts into formal representations that can be used for querying norms, compliance checking, and decision support. MIREL addresses both conceptual challenges, such as the role of legal interpretation in mining and reasoning, and computational challenges, such as the handling of big legal data, and the complexity of regulatory compliance. It bridges the gap between the community working on legal ontologies and NLP parsers and the community working on reasoning methods and formal logic. Moreover, it is the first project of its kind to involve industrial partners in the future development of innovative products and services in legal reasoning and their deployment in the market. In the context of the MIREL project, I co-supervise (together with Laura Alonso Alemany, Professor at the University of Cordoba, Argentina) two PhD students of the University of Cordoba: Cristian Cardellino (whom I supervised in his 6-months internship) and Milagro Teruel, both working on legal information extraction and argument mining applied to the legal domain.

In the context of normative reasoning and ethics, I am involved in a long term collaboration with Guido Governatori (Data61 - CSIRO, Australia) and Antonino Rotolo (Law Department, University of Bologna, Italy). This research is indirectly funded by the above cited projects. Always in this context, I supervised, together with Leendert van der Torre (University of Luxembourg) and Guido Governatori (Data61), the PhD of Javed Ahmed. The PhD was funded by the Joint International Doctoral (Ph.D.) Degree in Law, Science and Technology (LAST-JD) for the period: September 2013-2017. The title of the thesis is "Contextual

integrity and tie strength in online social networks: social theory, user study, ontology, and validation", and the PhD defense hold on September 29th, 2017 in Bologna.

Domain-independent projects

A new collaboration with the company Accenture is ongoing, and a project titled "Justifying Machine Decisions through Argumentation and Semantics" has been funded by the company through a 12-month research engineer position. The context of this project is the following. Robots helping humans in performing their everyday activities are becoming nowadays very popular, given the valuable impact they may bring on society, e.g., robots assisting elderly people in their places to support them in their everyday tasks. However, in order to concretely interact with humans, intelligent systems are required to show some human-like abilities such as the ability to explain their own decisions. The research question we target for this project is "how to explain and justify machine decisions to humans?". I will co-supervise the activity of this research engineer (Nicholas Halliwell), together with Freddy Lecue (Accenture), starting from July 2018.

This document is structured to describe my (unconventional) way in the broad area of argumentation theory, starting from formal computational models of argument to argumentation for humans. More precisely, the remainder of this document is organized as follows: Chapter 1 presents my contributions on modeling and decomposing abstract argumentation frameworks, Chapter 2 presents my contributions on reasoning about trust using argumentation theory, Chapter 3 presents my contributions on using argumentation theory and sources confidence to explain machine decisions, Chapter 4 presents my contributions on mining natural language arguments and their relations from texts, and finally, Chapter 5 presents my contributions on studying the relation between argumentation and persuasion, and emotions on debate participants. Conclusions summarize my work in the area of argumentation theory, and discuss future perspectives.

Chapter 2

Modularity and decomposability of argumentation frameworks

2.1 Introduction

This chapter synthesizes my contributions in the area of computational models of argument, dealing, more precisely, with the notion of modularity in abstract argumentation theory [131] and its computational properties. These contributions have been published in two papers in the Elsevier *Artificial Intelligence* journal [20, 301]:

- *Pietro Baroni, Guido Boella, Federico Cerutti, Massimiliano Giacomin, Leendert van der Torre, Serena Villata. On the Input/Output behavior of argumentation frameworks. Artif. Intell. 217: 144-197 (2014)*
- *Matthias Thimm, Serena Villata. The first international competition on computational models of argumentation: Results and analysis. Artif. Intell. 252: 267-294 (2017)*).

This chapter summarizes the results of my collaboration with Pietro Baroni and Massimiliano Giacomin from the University of Brescia, and the results of the first international competition on computational models of argumentation I organized with Matthias Thimm (Koblenz University). My collaboration with Leendert van der Torre (University of Luxembourg), my former PhD thesis supervisor, has continued since the defense of my thesis.

While the formalism of abstract argumentation frameworks [131] and the relevant argumentation semantics (see [13] for a survey) do not appear to have been designed with modularity in mind, investigating their relevant properties is an important research topic which, after having been somehow overlooked, is attracting increasing attention in recent years. An argumentation framework is basically a directed graph representing the conflicts between a set of arguments (the nodes of the graph) and an argumentation semantics can be regarded as a method to answer (typically in a non univocal way, i.e. producing a set of alternative answers) the "justification question": "Which is the justification status of arguments given the conflict?"

Referring to a representative set of semantics proposed in the literature, (namely admissible, complete, grounded, preferred, stable, semi-stable and ideal semantics), this chapter provides a systematic and comprehensive assessment of modularity in abstract argumentation, by identifying and analyzing in this context the formal counterparts of the general notions of *separability* and *interchangeability*.

2.2 Motivation and research questions

The "Merriam-Webster Learner's Dictionary" defines *modular* as "having parts that can be connected or combined in different ways" while the "Free Dictionary online" remarks that modularity is intended "for easy assembly and repair or flexible arrangement and use". As such, modularity is a highly desirable property, often enforced by design, in any kind of either material (like the popular Lego toys) or immaterial (like programs developed according to the object-oriented paradigm) artifacts, including knowledge representation and reasoning formalisms.

Roughly speaking, modularity involves two main properties, namely *separability* and *interchangeability* of modules. As to the former, it has to be possible to describe and analyse the global behavior of an artifact in terms of the combination of the local behaviors of the modules composing it. Each local behavior can be characterized individually in a way which is independent of the internal details of the other modules (and, in a sense, of the module itself) and captures only the connections and mutual interactions between the module and the other ones. To put it in other words, each module can be described as a black-box whose Input/Output behavior fully determines its role in the global behavior of any artifact it is plugged in. As to the latter, the interest in replacing a module with another one is very common and arises from a large variety of motivations, either at the operational or design level. Interchangeability of two modules requires first of all that they are compatible as far as the connections with the rest of the artifact are concerned, i.e. that the interfaces they expose are such that wherever one of the modules can be "plugged in", the other can too. Besides this *plug-level* interchangeability, it is of great interest to characterize the *behavior-level* interchangeability of modules, namely to identify the situations where internally different modules can be freely interchanged without affecting the global behavior of the artifact they belong to, since their Input/Output behavior is equivalent in this respect.

Given a partition of an argumentation framework into *partial* (or *local*) interacting subframeworks, analyzing separability consists in addressing the following issues:

- Is it possible to define a local counterpart of the notion of semantics? i.e. Is there a method to produce local answers to the justification question, taking into account the interactions with other subframeworks?
- Can the set of justification answers prescribed by the (global) semantics be obtained by properly combining (in a bottom-up fashion) the sets of local answers produced in the subframeworks by its local counterpart?
- symmetrically, Can the sets of local answers be obtained (in a top-down fashion) as projections onto the subframeworks of the global answers?

As to the first issue, we introduce the notion of *local function* for a subframework[1] and show that under very mild requirements, satisfied by all semantics considered in this chapter, it is possible (and easy) to identify the *canonical local function* for a global semantics. As to the second and third issues, we introduce the formal notions of top-down and bottom-up decomposability, which, jointly, correspond to the notion of (full) decomposability of an argumentation semantics.

Strong as it may seem, full decomposability with respect to every arbitrary partition of every argumentation framework is not unattainable. Indeed, we show that it is satisfied by some of the semantics considered

[1] Technically, a subframework is captured by the formal notion of *argumentation framework with input* provided in Definition 13.

in this chapter, while some others are able to achieve at least top-down decomposability and the remaining ones lack all decomposability properties.

As arbitrary partitions correspond to a completely free (if not anarchical) notion of modularity, we also consider a "tidier" style of partitioning, involving the graph-theoretical notion of *strongly connected components*. It turns out that, restricting the set of partitions this way, helps some, but not all, semantics to recover full decomposability.

Turning to interchangeability, we deal with both its plug-level and behavior-level aspects. As to the plug-level, borrowing some terminology from circuit theory, we introduce the notion of *argumentation multipole* as a generic replaceable argumentation component, namely a partial framework interacting through an input and output relation with an external set of invariant arguments.

Plug-level compatibility of two multipoles is a very relaxed notion, since it is only required that two multipoles refer to the same set of external arguments. This is motivated by the fact that imposing a tighter correspondence between Input/Ouput "terminals" of the multipoles would unnecessarily restrict the scope of the subsequent analysis on behavior-level compatibility. In fact, our analysis shows that a sensible notion of behavioral equivalence between multipoles (called *Input/Output equivalence*) can be introduced by requiring that the effect of the multipoles on the external arguments is the same: it may well be the case that multipoles with different "terminals" have the same effect in behavioral terms. Of course, Input/Output equivalence is a semantics-dependent notion since the behavior of a multipole can only be defined by referring to a specific semantics using the notion of local function mentioned above. In particular, it may be the case that two multipoles are equivalent with respect to some semantics and not equivalent with respect to another semantics.

Input/Output equivalence is the basis for the analysis of the operation of replacement within an argumentation framework. Basically, a replacement consists in substituting a part of the framework with a plug-level compatible multipole. While this notion *per se* allows for arbitrary substitutions, one is interested in analysing those replacements which have a sound basis. In this perspective, building on multipole equivalence, it is possible to identify the semantics-dependent notions of *legitimate* and *contextually legitimate* replacement. Briefly, a replacement is legitimate if it involves "fully" equivalent multipoles, while it is contextually legitimate if it involves multipoles which are equivalent in the context where the replacement takes place, while not necessarily being equivalent in other contexts. Cleary, legitimate replacements are a (typically strict) subset of contextually legitimate replacements.

One might expect that, given a semantics, legitimate (with respect to that semantics) replacements ensure that the invariant part of the framework is unaffected (in a sense, that it does not notice the change). This property is called *semantics transparency*. A stronger expectation (since the requirement on the replacements is weaker) would be that the invariant part of the framework is unaffected for any contextually legitimate replacement: this property is called *strong transparency*.

Natural as it may seem, transparency is not achieved by all semantics and requires a detailed analysis, showing that different levels of transparency are achieved by the semantics considered in this chapter, also taking into account different restrictions on the set of allowed replacements.

These results provide a reference context and fundamental answers to modularity-related issues in abstract argumentation, which, up to now, have been considered in the literature focusing on specific aspects and hence obtaining partial and problem-specific results. Moreover, while being theoretical by nature, the achievements of this chapter have several significant application-oriented implications.

On the one hand, semantics decomposability properties provide a sound basis for exploiting various forms of incremental computation which may deliver important efficiency gains in two main respects. First, they enable (and characterize the limits of) the application of divide-and-conquer strategies in the design of

algorithms for computational problems in abstract argumentation frameworks. As most of these problems are intractable in the worst case, facing reduced-size subproblems separately and then combining the partial results in an efficient manner may significantly improve performances on the average. Second, there is a significant application interest in argumentation dynamics, which captures all contexts where a given framework is updated incrementally, as a consequence of the acquisition of new information and/or of the actions of the participants to a multi-agent system. Clearly, if the modification to the initial framework is limited, one is interested to partially reuse the results of previous computations in the new framework rather than redoing all computations from scratch. Again, decomposability properties enable (and characterize the limits of) the use of incremental computation techniques based on the separation between modified and unmodified parts in the updated framework.

On the other hand, the notions and properties concerning multipole equivalence and semantics transparency are applicable in all contexts where there is an interest in replacing a part of a framework with another one. As an example, the activities of summarization and explanation involved in reasoning and communicating at different levels of granularity are, basically, alternative forms of replacement. In the former, a complex part of an argumentation process (e.g. the analysis and discussion of factual evidences in a legal case) is summarized (i.e. replaced) by a more synthetic representation (e.g. focusing on the facts which turn out to have an actual impact on the case decision) which, while leaving out unnecessary details, must ensure that the global outcome is preserved. Dually, explanation can be regarded as the replacement of a synthetic representation with a more detailed/articulated one, again ensuring that this does not induce undesired side-effects outside the replaced part. Further, and more specific of the abstract argumentation field, the basic formalism of argumentation frameworks is often used as a "ground level" representation for other richer and/or more specific formalisms. For instance, formalisms involving the explicit representation of preferences, values, and attacks to attacks can be translated (or flattened) to the basic formalism through suitable procedures. As these procedures typically consist of a set of local replacement rules, multipole equivalence and semantics transparency are very effective tools to analyze their behavior, soundness and applicability under various semantics.

The chapter is organized as follows. After recalling the necessary background in Section 2.3, the general notions concerning semantics decomposability are introduced and discussed in Section 2.4, while Section 2.5 provides decomposability results for the seven semantics considered in this chapter. Section 2.6 deals with the key technical notion of effect-dictated semantics and Section 2.7 then introduces the fundamental concepts concerning interchangeability, namely argumentation multipoles, their Input/Output equivalence, the replacement operator and the properties of semantics transparency. Section 2.8 analyzes the relationships between decomposability and transparency at a general level, while Section 2.9 provides transparency results for the seven semantics considered in this chapter. Application examples are given in Section 2.10, Section 2.11 discusses related works and, finally, Section 2.12 concludes the chapter.

2.3 Background

We follow the traditional definition of argumentation framework introduced by Dung [131] and define its restriction to a subset of arguments.

Definition 1. An *argumentation framework* is a pair $AF = (Ar, \rightarrow)$ in which Ar is a finite set of arguments and $\rightarrow \subseteq Ar \times Ar$. An argument A such that there is no B such that $(B,A) \in \rightarrow$ is called *initial*. An argument B such that $(B,B) \in \rightarrow$ is called *self-attacking*. Given a set $Args \subseteq Ar$, the restriction of AF to $Args$, denoted as $AF{\downarrow_{Args}}$ is the argumentation framework $(Args, \rightarrow \cap (Args \times Args))$.

Dung [131] presents several acceptability semantics which produce zero, one, or several sets of accepted arguments. These semantics are grounded on two main concepts, called conflict-freeness and defence.

Definition 2. (Conflict-free, Defence) Let $\Gamma \subseteq Ar$. A set Γ is *conflict-free* if and only if there exist no $A, B \in \Gamma$ such that $A \rightarrow B$. A set Γ *defends* an argument A if and only if for each argument $B \in Ar$ if B attacks A then there exists $C \in \Gamma$ such that C attacks B.

Definition 3. (Acceptability semantics) Let Γ be a conflict-free set of arguments, and let $\mathscr{D} : 2^{Ar} \mapsto 2^{Ar}$ be a function such that $\mathscr{D}(\Gamma) = \{A | \Gamma \, defends \, A\}$.

- Γ is *admissible* if and only if $\Gamma \subseteq \mathscr{D}(\Gamma)$.
- Γ is a *complete extension* if and only if $C = \mathscr{D}(\Gamma)$.
- Γ is a *grounded extension* if and only if it is the smallest (w.r.t. set inclusion) complete extension.
- Γ is a *preferred extension* if and only if it is a maximal (w.r.t. set inclusion) complete extension.
- Γ is a *stable extension* if and only if it is a preferred extension that attacks all arguments in $Ar \setminus \Gamma$.

The concepts of Dung's semantics are originally stated in terms of sets of arguments. It is equal to express these concepts using argument *labeling*. This approach has been proposed firstly by [176] and [305], and then further developed by [84] with the aim of providing quality postulates for dealing with the reinstatement of arguments. Given that $A \rightarrow B$ and $B \rightarrow C$, we have that argument A reinstates argument C, i.e., it makes argument C accepted by attacking the attacker of C.

In this chapter, we use the labelling-based approach to the definition of argumentation semantics. As shown in [81, 13], for the semantics considered in this chapter there is a direct correspondence with the "traditional" extension-based approach, hence the results presented in this chapter are valid in both approaches. The labelling-based definitions have been adopted only because they allow simpler proofs.

A labelling assigns to each argument of an argumentation framework a label taken from a predefined set Λ. For technical reasons, we define labellings both for argumentation frameworks and for arbitrary sets of arguments.

Definition 4. Let Λ be a set of labels. Given a set of arguments *Args*, a *labelling* of *Args* is a total function $Lab : Args \longrightarrow \Lambda$. The set of all *labellings* of *Args* is denoted as $\mathfrak{L}_{Args}$. Given an argumentation framework $AF = (Ar, \rightarrow)$, a *labelling* of AF is a labelling of Ar. The set of all *labellings* of AF is denoted as $\mathfrak{L}(AF)$. For a labelling *Lab* of *Args*, the restriction of *Lab* to a set of arguments $Args' \subseteq Args$, denoted as $Lab{\downarrow_{Args'}}$, is defined as $Lab \cap (Args' \times \Lambda)$.

We adopt the most common choice for Λ, i.e. $\{\texttt{in}, \texttt{out}, \texttt{undec}\}$, where the label `in` means that the argument is accepted, the label `out` means that the argument is rejected, and the label `undec` means that the status of the argument is undecided. As explained after Definition 10, an exception is made for stable semantics, which can be more conveniently defined assuming $\Lambda = \{\texttt{in}, \texttt{out}\}$. Given a labelling *Lab*, we write $\texttt{in}(Lab)$ for $\{A \mid Lab(A) = \texttt{in}\}$, $\texttt{out}(Lab)$ for $\{A \mid Lab(A) = \texttt{out}\}$ and $\texttt{undec}(Lab)$ for $\{A \mid Lab(A) = \texttt{undec}\}$.

A labelling-based semantics prescribes a set of labellings for each argumentation framework.

Definition 5. Given an argumentation framework $AF = (Ar, \rightarrow)$, a labelling-based semantics $\mathbf{S}$ associates with AF a subset of $\mathfrak{L}(AF)$, denoted as $\mathbf{L_S}(AF)$.

In general, a semantics encompasses a set of alternative labellings for a single argumentation framework. However, a semantics may be defined so that a unique labelling is always prescribed, i.e. for every argumentation framework AF, $|\mathbf{L_S}(AF)| = 1$. In this case the semantics is said to be *single-status*, while in the general case it is said to be *multiple-status*.

In the labelling-based approach, a semantics definition relies on some *legality* constraints relating the label of an argument to those of its attackers.

Definition 6. Let *Lab* be a labelling of the argumentation framework $(Ar, \rightarrow)$. An `in`-labelled argument is said to be *legally* `in` iff all its attackers are labelled `out`. An `out`-labelled argument is said to be *legally* `out` iff it has at least one attacker that is labelled `in`. An `undec`-labelled argument is said to be *legally* `undec` iff not all its attackers are labelled `out` and it does not have an attacker that is labelled `in`.

We now introduce the definitions of labellings corresponding to traditional admissible[2] and complete semantics.

Definition 7. Let $AF = (Ar, \rightarrow)$ be an argumentation framework. An *admissible labelling* is a labelling *Lab* where every `in`-labelled argument is legally `in` and every `out`-labelled argument is legally `out`.

Definition 8. A *complete labelling* is a labelling where every `in`-labelled argument is legally `in`, every `out`-labelled argument is legally `out` and every `undec`-labelled argument is legally `undec`.

On this basis, the labelling-based definitions of several argumentation semantics can be introduced. To simplify the technical treatment in the following, grounded and preferred semantics are defined by referring to the commitment relation between labellings [13].

Definition 9. Let Lab_1 and Lab_2 be two labellings. We say that Lab_2 is *more or equally committed* than Lab_1 ($Lab_1 \sqsubseteq Lab_2$) iff $\mathtt{in}(Lab_1) \subseteq \mathtt{in}(Lab_2)$ and $\mathtt{out}(Lab_1) \subseteq \mathtt{out}(Lab_2)$.

Definition 10. Let $AF = (Ar, \rightarrow)$ be an argumentation framework. A *stable labelling* of AF is a complete labelling without `undec`-labelled arguments. The *grounded labelling* of AF is the minimal (w.r.t. $\sqsubseteq$) labelling among all complete labellings. A *preferred labelling* of AF is a maximal (w.r.t. $\sqsubseteq$) labelling among all complete labellings. The *ideal labelling* of AF is the maximal (under $\sqsubseteq$) complete[3] labelling *Lab* that is less or equally committed than each preferred labelling of AF (i.e. for each preferred labelling Lab_P it holds that $Lab \sqsubseteq Lab_P$). A *semi-stable* labelling of AF is a complete labelling *Lab* where $\mathtt{undec}(Lab)$ is minimal (w.r.t. set inclusion) among all complete labellings.

While stable semantics is defined by assuming $\Lambda = \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}$, the definition of stable labelling entails that stable semantics can be equivalently defined with reference to the set of labels $\Lambda = \{\mathtt{in}, \mathtt{out}\}$. In this case, a stable labelling is simply a complete labelling, since the codomain Λ does not include `undec`. In the sequel we implicitly assume that, for stable semantics only, $\Lambda = \{\mathtt{in}, \mathtt{out}\}$: this allows a simpler treatment of such semantics without any loss of generality.

The uniqueness of the grounded and the ideal labelling has been proved in [82]. Accordingly, grounded and ideal semantics are single-status, the other semantics are multiple-status. Admissible, complete, stable,

[2] It can be remarked that (unlike the other semantics) admissible labellings are not in a one-to-one correspondence to admissible sets since several admissible labellings might correspond to the same admissible set.

[3] Literally, the original definition refers to an admissible labelling rather than a complete labelling. However, the definition adopted here is equivalent to the original one, since it can be shown that the ideal labelling is a complete labelling [82].

grounded, preferred, ideal and semi-stable semantics are denoted in the following as **AD**, **CO**, **ST**, **GR**, **PR**, **ID** and **SST**, respectively.

We also recall the traditional notions of skeptical and credulous justification of an argument with respect to a semantics.

Definition 11. Given a labelling-based semantics **S** and an argumentation framework AF, an argument A is *skeptically justified* under **S** if $\forall Lab \in \mathbf{L_S}(AF)\ Lab(A) = \texttt{in}$; an argument A is *credulously justified* under **S** if $\exists Lab \in \mathbf{L_S}(AF) : Lab(A) = \texttt{in}$.

Finally, a comment is in order on a special case of argumentation framework that is explicitly considered in the chapter, i.e. the empty argumentation framework $AF_\emptyset \triangleq (\emptyset,\emptyset)$. By definition the only possible labelling of $AF_\emptyset$ is the empty set, thus a semantics can either prescribe $\emptyset$ for $AF_\emptyset$ or it can prescribe no labelling at all. In this respect, for any semantics **S** introduced above it holds $\mathbf{L_S}(AF_\emptyset) = \{\emptyset\}$, i.e. the empty set is actually prescribed by **S**. Note in particular that $\emptyset$ is a stable labelling, since it is complete and does not include `undec`-labelled arguments.

2.4 Decomposability of Argumentation Semantics

The notion of local function

The first step to define the notion of semantics decomposability is to introduce a formal setting to express the interactions between the partial frameworks induced by an arbitrary partitioning of an argumentation framework. Intuitively, given an argumentation framework $AF = (Ar, \rightarrow)$ and a subset $Args$ of its arguments, the elements affecting $AF{\downarrow_{Args}}$ include the arguments attacking $Args$ from the outside, called *input* arguments, and the attack relation from the input arguments to $Args$, called *conditioning relation*.

Definition 12. Given $AF = (Ar, \rightarrow)$ and a set $Args \subseteq Ar$, the *input* of $Args$, denoted as $Args^{\text{inp}}$, is the set $\{B \in Ar \setminus Args \mid \exists A \in Args, (B,A) \in \rightarrow\}$, the *conditioning relation* of $Args$, denoted as $Args^R$, is defined as $\rightarrow \cap (Args^{\text{inp}} \times Args)$.

Example 1. Consider $AF = (\{A,B,C,D\}, \{(A,B),(B,C),(C,A),(A,D),(D,A)\})$ with reference to the partial frameworks induced by the sets $\{A,B,C\}$ and $\{D\}$ (see Figure 2.1). It holds that $\{A,B,C\}^{\text{inp}} = \{D\}$ and $\{A,B,C\}^R = \{(D,A)\}$, while $\{D\}^{\text{inp}} = \{A\}$ and $\{D\}^R = \{(A,D)\}$.

Given a partial argumentation framework $AF{\downarrow_{Args}}$ (possibly AF itself) affected by a (possibly empty) set of arguments $Args^{\text{inp}}$ attacking $Args$ according to $Args^R$, one may wonder whether fixing the labelling assigned to the input arguments allows one to determine the set of labellings of $AF{\downarrow_{Args}}$. As shown in the following, this question cannot be answered once and for all, since different semantics exhibit different behaviours in this respect, and, for some semantics, a dependency holds under specific constraints on the considered partition of the argumentation framework. In order to express such a dependency (whenever it holds), we introduce the notions of *argumentation framework with input*, consisting of an argumentation framework $AF = (Ar, \rightarrow)$ (playing the role of a partial argumentation framework), a set of external input arguments $\mathscr{I}$, a labelling $L_\mathscr{I}$ assigned to them and an attack relation $R_\mathscr{I}$ from $\mathscr{I}$ to Ar, and of a *local function* which, given an argumentation framework with input, returns a corresponding set of labellings of AF.

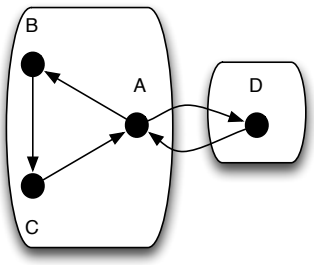


Figure 2.1: Running example: a partition of a simple framework (Examples 1 - 5).

Definition 13. An *argumentation framework with input* is a tuple $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, including an argumentation framework[4] $AF = (Ar, \rightarrow)$, a set of arguments $\mathscr{I}$ such that $\mathscr{I} \cap Ar = \emptyset$, a labelling $L_{\mathscr{I}} \in \mathfrak{L}_{\mathscr{I}}$ and a relation $R_{\mathscr{I}} \subseteq \mathscr{I} \times Ar$. A *local function* assigns to any argumentation framework with input a (possibly empty) set of labellings of AF, i.e. $F(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}}) \in 2^{\mathfrak{L}(AF)}$.

For any semantics, a "sensible" local function, called *canonical local function*, is the one that describes the labellings of the so-called standard argumentation frameworks.

Definition 14. Given an argumentation framework with input $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, the standard argumentation framework w.r.t. $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ is defined as $AF' = (Ar \cup \mathscr{I}', \rightarrow \cup R'_{\mathscr{I}})$, where $\mathscr{I}' = \mathscr{I} \cup \{A' \mid A \in \mathtt{out}(L_{\mathscr{I}})\}$ and $R'_{\mathscr{I}} = R_{\mathscr{I}} \cup \{(A', A) \mid A \in \mathtt{out}(L_{\mathscr{I}})\} \cup \{(A, A) \mid A \in \mathtt{undec}(L_{\mathscr{I}})\}$.

Roughly, the standard argumentation framework puts AF under the influence of $(\mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, by adding $\mathscr{I}$ to Ar and $R_{\mathscr{I}}$ to $\rightarrow$, and by enforcing[5] the label $L_{\mathscr{I}}$ for the arguments of $\mathscr{I}$ in this way:

- for each argument $A \in \mathscr{I}$ such that $L_{\mathscr{I}}(A) = \mathtt{out}$, an unattacked argument A' is included which attacks A, in order to get A labelled `out` by all labellings of AF';
- for each argument $A \in \mathscr{I}$ such that $L_{\mathscr{I}}(A) = \mathtt{undec}$, a self-attack is added to A in order to get it labelled `undec` by all labellings of AF';
- each argument $A \in \mathscr{I}$ such that $L_{\mathscr{I}}(A) = \mathtt{in}$ is left unattacked, so that it is labelled `in` by all labellings of AF'.

Definition 15. Given a semantics $\mathbf{S}$, the *canonical local function* of $\mathbf{S}$ (also called local function of $\mathbf{S}$) is defined as $F_{\mathbf{S}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}}) = \{Lab{\downarrow_{Ar}} \mid Lab \in \mathbf{L}_{\mathbf{S}}(AF')\}$, where $AF = (Ar, \rightarrow)$ and AF' is the standard argumentation framework w.r.t. $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$.

Note that in the case of stable semantics *undec* $\notin \Lambda$, thus $R'_{\mathscr{I}}$ does not include self-attacks.

In case $\mathscr{I} = \emptyset$ (entailing $L_{\mathscr{I}} = \emptyset$ and $R_{\mathscr{I}} = \emptyset$) the canonical local function returns the labellings of AF, as shown by Proposition 1. All proofs are available online at `https://www.sciencedirect.com/science/article/pii/S0004370214001131?via%3Dihub`.

[4] In the following, unless otherwise specified, we will implicitly assume $AF = (Ar, \rightarrow)$.

[5] Actually, the enforcement is a bit different for admissible semantics. This exception has no consequences on the technical development of the chapter.

Proposition 1. Given a semantics **S** and an argumentation framework AF, $F_{\mathbf{S}}(AF,\emptyset,\emptyset,\emptyset) = \mathbf{L_S}(AF)$.

While the canonical local function is defined for any semantics, its definition is best suited for *complete-compatible* semantics, i.e. semantics satisfying a number of intuitive constraints.

Definition 16. A semantics **S** is *complete-compatible* iff the following conditions hold:

1. For any argumentation framework $AF = (Ar, \rightarrow)$, every labelling $\mathbf{L} \in \mathbf{L_S}(AF)$ satisfies the following conditions:
   - if $A \in Ar$ is initial, then $\mathbf{L}(A) = \texttt{in}$
   - if $B \in Ar$ and there is an initial argument A which attacks B, then $\mathbf{L}(B) = \texttt{out}$
   - if $C \in Ar$ is self-attacking, and there are no attackers of C besides C itself, then $\mathbf{L}(C) = \texttt{undec}$
2. for any set of arguments $\mathscr{I}$ and any labelling $L_{\mathscr{I}} \in \mathfrak{L}_{\mathscr{I}}$, the argumentation framework $AF' = (\mathscr{I}', \rightarrow')$, where $\mathscr{I}' = \mathscr{I} \cup \{A' \mid A \in \texttt{out}(L_{\mathscr{I}})\}$ and $\rightarrow' = \{(A', A) \mid A \in \texttt{out}(L_{\mathscr{I}})\} \cup \{(A, A) \mid A \in \texttt{undec}(L_{\mathscr{I}})\}$, admits a (unique) labelling, i.e. $|\mathbf{L_S}(AF')| = 1$.

It should be noted that, in case $\texttt{undec} \notin \Lambda$, the third bullet of condition 1 entails that there is no labelling if a self-attacking argument C is attacked by C only, and in condition 2 it necessarily holds that $\texttt{undec}(L_{\mathscr{I}}) = \emptyset$.

As shown by Proposition 2, the requirements of the previous definition guarantee that the construction of the standard argumentation framework makes sense, i.e. given a standard argumentation framework w.r.t. $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, a complete-compatible semantics enforces the labelling $L_{\mathscr{I}}$ for the arguments of $\mathscr{I}$ as described above.

Proposition 2. Let **S** be a complete-compatible semantics and let $AF' = (Ar \cup \mathscr{I}', \rightarrow \cup R'_{\mathscr{I}})$ be the standard argumentation framework w.r.t. an argumentation framework with input $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$. Then for any $Lab \in \mathbf{L_S}(AF')$ it holds that $Lab{\downarrow}_{\mathscr{I}'} = \{(A', \texttt{in}) \mid A \in \texttt{out}(L_{\mathscr{I}})\} \cup L_{\mathscr{I}}$ and $Lab{\downarrow}_{\mathscr{I}} = L_{\mathscr{I}}$.

Moreover, when applied to the empty argumentation framework (which by definition does not receive attacks from $\mathscr{I}$) the canonical local function of a complete-compatible semantics always returns the empty set as a unique labelling.

Proposition 3. Given a complete-compatible semantics **S**, a set of arguments $\mathscr{I}$ and a labelling $L_{\mathscr{I}} \in \mathfrak{L}_{\mathscr{I}}$, it holds that $F_{\mathbf{S}}(AF_{\emptyset}, \mathscr{I}, L_{\mathscr{I}}, \emptyset) = \{\emptyset\}$.

Taking into account Proposition 1 this result entails that $\mathbf{L_S}(AF_{\emptyset}) = \{\emptyset\}$, corresponding to the second requirement of Definition 16 with $\mathscr{I} = \emptyset$.

All the semantics considered in the chapter are complete-compatible, with the exception of admissible semantics.

Proposition 4. **GR**, **CO**, **ST**, **PR**, **SST**, **ID** are all complete-compatible semantics.

Admissible semantics is not complete-compatible, as it can be seen by considering e.g. the argumentation framework $AF = (\{A\}, \emptyset)$, where $\mathbf{L_{AD}}(AF) = \{(A, \texttt{undec}), (A, \texttt{in})\}$.

The following example clarifies the notion of canonical local function, considering in particular complete semantics.

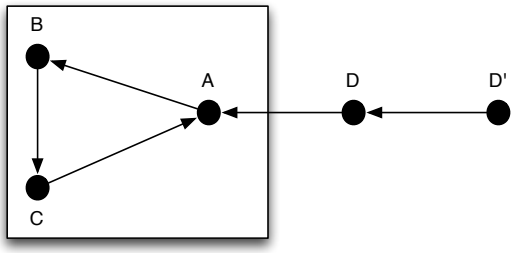


Figure 2.2: The standard argumentation framework w.r.t. $(AF{\downarrow}_{\{A,B,C\}}, \{D\}, \{(D,\texttt{out})\}, \{(D,A)\})$ (Example 2).

Example 2. Let us refer again to the argumentation framework AF of Figure 2.1. For the canonical local function of complete semantics it holds that $F_{\mathbf{CO}}(AF{\downarrow}_{\{A,B,C\}}, \{D\}, \{(D,\texttt{out})\}, \{(D,A)\}) = \{\{(A,\texttt{undec}),(B,\texttt{undec}),(C,\texttt{undec})\}\}$, due to the fact that the standard argumentation framework w.r.t. $(AF{\downarrow}_{\{A,B,C\}}, \{D\}, \{(D,\texttt{out})\}, \{(D,A)\})$, shown in Figure 2.2, admits as the unique complete labelling $\{(D',\texttt{in}),(D,\texttt{out}),(A,\texttt{undec}),(B,\texttt{undec}),(C,\texttt{undec})\}$. In a similar way, it is easy to show that $F_{\mathbf{CO}}(AF{\downarrow}_{\{A,B,C\}}, \{D\}, \{(D,\texttt{in})\}, \{(D,A)\}) = \{\{(A,\texttt{out}),(B,\texttt{in}),(C,\texttt{out})\}\}$ and $F_{\mathbf{CO}}(AF{\downarrow}_{\{A,B,C\}}, \{D\}, \{(D,\texttt{undec})\}, \{(D,A)\}) = \{\{(A,\texttt{undec}),(B,\texttt{undec}),(C,\texttt{undec})\}\}$.
Considering the application of $F_{\mathbf{CO}}$ to $AF{\downarrow}_{\{D\}}$, $F_{\mathbf{CO}}(AF{\downarrow}_{\{D\}}, \{A\}, \{(A,\texttt{out})\}, \{(A,D)\}) = \{\{(D,\texttt{in})\}\}$, $F_{\mathbf{CO}}(AF{\downarrow}_{\{D\}}, \{A\}, \{(A,\texttt{in})\}, \{(A,D)\}) = \{\{(D,\texttt{out})\}\}$ and $F_{\mathbf{CO}}(AF{\downarrow}_{\{D\}}, \{A\}, \{(A,\texttt{undec})\}, \{(A,D)\}) = \{\{(D,\texttt{undec})\}\}$.

As shown in Section 2.5, for any semantics considered in this chapter the local function admits a compact representation, without the need to refer to standard argumentation frameworks.

Decomposability properties of argumentation semantics

We now aim at introducing a formal notion of semantics decomposability. To this purpose, consider a generic argumentation framework $AF = (Ar, \rightarrow)$ and an arbitrary partition of Ar, i.e. a set $\{P_1, \ldots, P_n\}$ such that $\forall i \in \{1, \ldots, n\}$ $P_i \subseteq Ar$ and $P_i \neq \emptyset$, $\bigcup_{i=1\ldots n} P_i = Ar$ and $P_i \cap P_j = \emptyset$ for $i \neq j$. Such a partition identifies the restricted argumentation frameworks $AF{\downarrow}_{P_1}, \ldots, AF{\downarrow}_{P_n}$, that affect each other with the relevant input arguments and conditioning relations as stated in Definition 12. Intuititively a semantics $\mathbf{S}$ is decomposable if $\mathbf{S}$ can be put in correspondence with a local function F such that:

- every labelling prescribed by $\mathbf{S}$ on AF, namely every element of $\mathbf{L}_{\mathbf{S}}(AF)$, corresponds to the union of n "compatible" labellings $L_{P_1}, \ldots, L_{P_n}$ of the restricted argumentation frameworks, all of them obtained applying F;
- in turn, each union of n "compatible" labellings $L_{P_1}, \ldots, L_{P_n}$ obtained applying F to the restricted frameworks gives rise to a labelling of AF.

The "compatibility" constraint mentioned above reflects the fact that any labelling of a restricted framework is used by F for computing the other ones: L_{P_i} plays a role in determining $L_{P_1}, \ldots, L_{P_{i-1}}, L_{P_{i+1}}, \ldots, L_{P_n}$ and vice versa. This means that $L_{P_1}, \ldots, L_{P_n}$ are "compatible" if each L_{P_i} is produced by F for $AF{\downarrow}_{P_i}$ with the

input arguments P_i^{inp} labelled according to $L_{P_1},\ldots,L_{P_{i-1}},L_{P_{i+1}},\ldots,L_{P_n}$. Definition 17 synthesizes all these considerations.

Definition 17. A semantics **S** is *fully decomposable* (or simply *decomposable*) iff there is a local function F such that for every argumentation framework $AF=(Ar,\rightarrow)$ and every partition $\mathscr{P}=\{P_1,\ldots,P_n\}$ of Ar, $\mathbf{L_S}(AF) = \mathscr{U}(\mathscr{P},AF,F)$ where $\mathscr{U}(\mathscr{P},AF,F) \triangleq \{L_{P_1} \cup \ldots \cup L_{P_n} \mid L_{P_i} \in F(AF{\downarrow_{P_i}}, P_i^{\text{inp}}, (\bigcup_{j=1\ldots n, j\neq i} L_{P_j}){\downarrow_{P_i^{\text{inp}}}}, P_i^R)\}$.

Example 3. Considering again the argumentation framework AF of Figure 2.1 and the partition $\{\{A,B,C\}, \{D\}\}$, full decomposability of complete semantics requires a local function such that the labellings of AF are exactly those obtained by the union of the compatible labellings of $AF{\downarrow_{\{A,B,C\}}}$ and $AF{\downarrow_{\{D\}}}$ given by the local function itself. Let us consider the canonical local function[6] of **CO** (refer to Example 2). The labelling $\{(A,\texttt{out}),(B,\texttt{in}),(C,\texttt{out})\}$ is compatible with $\{(D,\texttt{in})\}$, since the first is obtained by $F_{\mathbf{CO}}$ with D labelled `in`, and the latter is obtained by $F_{\mathbf{CO}}$ with A labelled `out`. On the other hand, the labelling $\{(A,\texttt{out}),(B,\texttt{in}),(C,\texttt{out})\}$ is not compatible e.g. with $\{(D,\texttt{out})\}$. Overall, exactly two global labellings arise from the combinations of the compatible outcomes of $F_{\mathbf{CO}}$, namely $\{(A,\texttt{undec}),(B,\texttt{undec}),(C,\texttt{undec}),(D,\texttt{undec})\}$ and $\{(A,\texttt{out}),(B,\texttt{in}),(C,\texttt{out}),(D,\texttt{in})\}$, corresponding to the complete labellings of AF.

The behavior of complete semantics in this example is not incidental: we will prove in Section 2.5 that complete semantics is fully decomposable.

Proposition 5 shows that, if a complete-compatible semantics **S** is fully decomposable, then the local function appearing in Definition 17 coincides with the canonical local function $F_{\mathbf{S}}$.

Proposition 5. Given a complete-compatible semantics **S**, if **S** is fully decomposable then there is a unique local function satisfying the conditions of Definition 17, coinciding with the canonical local function $F_{\mathbf{S}}$.

Full decomposability can be viewed as the conjunction of two partial decomposability properties, namely *top-down decomposability* and *bottom-up* decomposability.

In words, a semantics is top-down decomposable if the procedure to compute the global labellings identified by Definition 17 is complete, i.e. all of the global labellings can be obtained by combining the labellings prescribed by $F_{\mathbf{S}}$ for the restricted subframeworks, even if putting together labellings of the restricted subframeworks may give rise to some "spurious" labellings besides the correct ones. The following definition formalizes this intuition.

Definition 18. A complete-compatible semantics **S** is *top-down decomposable* iff for any argumentation framework $AF=(Ar,\rightarrow)$ and any partition $\mathscr{P}=\{P_1,\ldots,P_n\}$ of Ar, it holds that $\mathbf{L_S}(AF) \subseteq \mathscr{U}(\mathscr{P},AF,F_{\mathbf{S}})$.

While top-down decomposability corresponds to completeness of the procedure identified by Definition 17, bottom-up decomposability requires its soundness, i.e. that any combination of local labellings is a global labelling, while it is not guaranteed that all global labellings can be obtained in this way.

Definition 19. A complete-compatible semantics **S** is *bottom-up decomposable* iff for any argumentation framework $AF=(Ar,\rightarrow)$ and any partition $\mathscr{P}=\{P_1,\ldots,P_n\}$ of Ar, it holds that $\mathbf{L_S}(AF) \supseteq \mathscr{U}(\mathscr{P},AF,F_{\mathbf{S}})$.

[6] It is shown in Proposition 5 that considering the canonical local function is without loss of generality.

A comment on the two definitions above is in order. While the definition of full decomposability applies to any kind of semantics and requires the existence of a local function satisfying the decomposability property, Definitions 18 and 19 are restricted to complete-compatible semantics and refer to the canonical local function $F_{\mathbf{S}}$ to avoid triviality: the local function returning all the possible labellings of AF trivially satisfies the inclusion condition of Definition 18 for any semantics, while the local function always returning the empty set trivially satisfies the condition of Definition 19. This is the reason why both definitions refer to the specific canonical local function, which makes sense for complete-compatible semantics in the light of Proposition 5. If a semantics is not complete-compatible[7] then the notion of canonical local function is meaningless, since the labelling $L_{\mathscr{I}}$ would not be in general enforced for the arguments of $\mathscr{I}$ in the standard argumentation framework w.r.t. $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ (see Proposition 2).

As shown in Section 2.5, some semantics that do not satisfy full decomposability are still able to satisfy top-down decomposability. Moreover, there are semantics that do not satisfy either of them: in this case it is interesting to investigate whether decomposability holds by restricting the possible partitions of the argumentation frameworks to those satisfying a given set of constraints. To express this restriction, we first introduce the notion of partition selector.

Definition 20. A *partition selector* $\mathscr{F}$ is a function receiving as input an argumentation framework $AF = (Ar, \rightarrow)$ and returning a set of partitions of Ar.

A partition selector is defined as a function of argumentation frameworks, since different argumentation frameworks with the same set of arguments may allow different sets of partitions, depending on the attack relation.

The decomposability notions introduced so far can then be extended to take into account a specific restriction on the considered partitions.

Definition 21. Let $\mathscr{F}$ be a partition selector. A complete-compatible semantics $\mathbf{S}$ is *top-down decomposable w.r.t.* $\mathscr{F}$ iff for any argumentation framework AF and any partition $\mathscr{P} = \{P_1, \ldots, P_n\} \in \mathscr{F}(AF)$, it holds that $\mathbf{L}_{\mathbf{S}}(AF) \subseteq \mathscr{U}(\mathscr{P}, AF, F_{\mathbf{S}})$. A complete-compatible semantics $\mathbf{S}$ is *bottom-up decomposable w.r.t.* $\mathscr{F}$ iff for any argumentation framework AF and any partition $\{P_1, \ldots, P_n\} \in \mathscr{F}(AF)$, $\mathbf{L}_{\mathbf{S}}(AF) \supseteq \mathscr{U}(\mathscr{P}, AF, F_{\mathbf{S}})$. A complete-compatible semantics is *fully decomposable* (or simply *decomposable*) *w.r.t. a partition selector* $\mathscr{F}$ iff it is both top-down and bottom-up decomposable w.r.t. $\mathscr{F}$.

Of course, full decomposability, top-down decomposability and bottom-up decomposability as introduced in Definitions 17, 18 and 19, respectively, are equivalent to the corresponding decomposability properties w.r.t. $\mathscr{F}_{\mathtt{ALL}}$, i.e. the selector returning all possible partitions.

Definition 22. For any argumentation framework $AF = (Ar, \rightarrow)$, $\mathscr{F}_{\mathtt{ALL}}(AF) \triangleq \{\{P_1, \ldots, P_n\} \mid \{P_1, \ldots, P_n\}$ is a partition of $Ar\}$.

Apart from this limit case, a particular partition selector that has received attention in the literature and will be considered in this chapter is the one based on the notion of strongly connected component (SCC) of an argumentation framework. Its importance is due to the fact that most argumentation semantics in the literature are *SCC-recursive* [245], which, briefly, means that the semantics can be defined in terms of a *base function* operating at the level of single strongly connected components. Roughly, this also implies

[7] Besides admissible semantics, in the literature there are a few examples of non complete-compatible semantics, like stage semantics [306] and various forms of prudent semantics [115].

	AD	CO	ST	GR	PR	ID	SST
Full decomposability (Def. 17)	Yes	Yes	Yes	No	No	No	No
Top-down decomposability (Def. 18)	-	Yes	Yes	Yes	Yes	No	No
Bottom-up decomposability (Def. 19)	-	Yes	Yes	No	No	No	No
Full decomposability w.r.t. $\mathscr{F}_{\cup\mathsf{SCC}}$ and $\mathscr{F}_{\mathsf{SCC}}$ (Def. 21)	Yes	Yes	Yes	Yes	Yes	No	No
Top-down decomposability w.r.t. $\mathscr{F}_{\cup\mathsf{SCC}}$ and $\mathscr{F}_{\mathsf{SCC}}$ (Def. 21)	-	Yes	Yes	Yes	Yes	No	No
Bottom-up decomposability w.r.t. $\mathscr{F}_{\cup\mathsf{SCC}}$ and $\mathscr{F}_{\mathsf{SCC}}$ (Def. 21)	-	Yes	Yes	Yes	Yes	No	No

Table 2.1: Decomposability properties of argumentation semantics.

that an incremental computation procedure based on the decomposition of the framework into its strongly connected components can be defined, a property exploited in several subsequent works [201, 272, 98]. Here we introduce the necessary basic definitions, leaving further discussion on this subject to Section 2.11.

Definition 23. Given an argumentation framework $AF = (Ar, \rightarrow)$, the set of strongly connected components of AF, denoted as SCCS_{AF}, consists of the equivalence classes of arguments induced by the binary relation of path-equivalence, i.e. the relation $\rho(A,B)$ defined over $Ar \times Ar$ such that $\rho(A,B)$ holds if and only if $A = B$ or there are directed paths from A to B *and* from B to A in AF.

For instance, the argumentation framework of Figure 2.1 has a unique strongly connected component including all of the arguments, while for the argumentation framework AF of Figure 2.2 it holds that $\text{SCCS}_{AF} = \{\{D'\},\{D\},\{A,B,C\}\}$.

At least two partition selectors based on strongly connected components can be considered. The simplest selector, denoted as $\mathscr{F}_{\mathsf{SCC}}$, includes for each argumentation framework AF the unique partition consisting of the strongly connected components SCCS_{AF}. A second selector, denoted as $\mathscr{F}_{\cup\mathsf{SCC}}$, includes all the partitions such that every element is the union of some (possibly unconnected) strongly connected components.

Definition 24. For any argumentation framework $AF = (Ar, \rightarrow)$, $\mathscr{F}_{\mathsf{SCC}}(AF) \triangleq \{\text{SCCS}_{AF}\} \setminus \{\emptyset\}$, $\mathscr{F}_{\cup\mathsf{SCC}}(AF) \triangleq \{\{P_1,\ldots,P_n\} \mid \{P_1,\ldots,P_n\}$ is a partition of Ar and $\forall i\ ((S \in \text{SCCS}_{AF} \wedge P_i \cap S \neq \emptyset) \rightarrow S \subseteq P_i)\}$.

It is immediate to see that, for any AF, $\mathscr{F}_{\mathsf{SCC}}(AF) \subseteq \mathscr{F}_{\cup\mathsf{SCC}}(AF)$. As to the first part of the definition, note that the set SCCS_{AF} includes $\emptyset$ only in case $AF = AF_{\emptyset}$, which does not admit any partition (since all the elements of a partition must be nonempty), thus $\mathscr{F}_{\mathsf{SCC}}(AF_{\emptyset}) = \emptyset$.

2.5 Analyzing semantics decomposability

In this section we discuss the decomposability properties of the semantics reviewed in Section 2.3. A synthetic view of the results is given in Table 2.1 (note that for all semantics full, top-down and bottom-up decomposability w.r.t. $\mathscr{F}_{\cup\mathsf{SCC}}$ turn out to be satisfied if and only if full, top-down and bottom-up decomposability w.r.t. $\mathscr{F}_{\mathsf{SCC}}$ are satisfied, respectively). Since admissible semantics is not complete-compatible, only the notion of full decomposability is applicable to it.

Admissible and complete semantics

We first analyze admissible and complete semantics, since they are the basis for the other ones considered in this chapter: according to Definition 10, stable, grounded, preferred, ideal, and semi-stable semantics

select labellings among the complete ones, which are admissible by definition. Given this, it would be very unpleasant if complete (and thus admissible) semantics would not be decomposable. As shown by Theorems 1 and 3, luckily both admissible and complete semantics turn out to be fully decomposable.

The following definition introduces the canonical local function of admissible semantics, by extending the definition of admissible labelling in order to account for "external" input arguments in the obvious way. The proof that the definition is correct is provided by Theorem 2.

Definition 25. Given an argumentation framework with input $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, $F_{\mathbf{AD}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}}) \triangleq$
$\{Lab \in \mathfrak{L}(AF) \mid$
$Lab(A) = \mathtt{in} \rightarrow ((\forall B \in Ar : (B,A) \in \rightarrow, Lab(B) = \mathtt{out}) \wedge (\forall B \in \mathscr{I} : (B,A) \in R_{\mathscr{I}}, L_{\mathscr{I}}(B) = \mathtt{out}))$,
$Lab(A) = \mathtt{out} \rightarrow ((\exists B \in Ar : (B,A) \in \rightarrow \wedge Lab(B) = \mathtt{in}) \vee (\exists B \in \mathscr{I} : (B,A) \in R_{\mathscr{I}} \wedge L_{\mathscr{I}}(B) = \mathtt{in}))\}$.

Theorem 1 proves that admissible semantics is fully decomposable, showing that the local function $F_{\mathbf{AD}}$ introduced in Definition 25 satisfies the conditions of Definition 17.

Theorem 1. *Admissible semantics* **AD** *is fully decomposable, with* $F_{\mathbf{AD}}$ *satisfying the conditions of Definition 17.*

The following theorem confirms that Definition 25 actually corresponds to the canonical local function of admissible semantics.

Theorem 2. *The canonical local function of admissible semantics is* $F_{\mathbf{AD}}$*, as defined in Definition 25.*

Also the canonical local function of complete semantics can be guessed on the basis of the definition of complete labelling.

Definition 26. Given an argumentation framework with input $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, $F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}}) \triangleq$
$\{Lab \in \mathfrak{L}(AF) \mid$
$Lab(A) = \mathtt{in} \rightarrow ((\forall B \in Ar : (B,A) \in \rightarrow, Lab(B) = \mathtt{out}) \wedge (\forall B \in \mathscr{I} : (B,A) \in R_{\mathscr{I}}, L_{\mathscr{I}}(B) = \mathtt{out}))$,
$Lab(A) = \mathtt{out} \rightarrow ((\exists B \in Ar : (B,A) \in \rightarrow \wedge Lab(B) = \mathtt{in}) \vee (\exists B \in \mathscr{I} : (B,A) \in R_{\mathscr{I}} \wedge L_{\mathscr{I}}(B) = \mathtt{in}))$,
$Lab(A) = \mathtt{undec} \rightarrow (((\forall B \in Ar : (B,A) \in \rightarrow, Lab(B) \neq \mathtt{in}) \wedge (\forall B \in \mathscr{I} : (B,A) \in R_{\mathscr{I}}, L_{\mathscr{I}}(B) \neq \mathtt{in})) \wedge ((\exists B \in Ar : (B,A) \in \rightarrow \wedge Lab(B) = \mathtt{undec}) \vee (\exists B \in \mathscr{I} : (B,A) \in R_{\mathscr{I}} \wedge L_{\mathscr{I}}(B) = \mathtt{undec})))\}$.

It is easy to see that $F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}}) \subseteq F_{\mathbf{AD}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, i.e. every "locally complete" labelling is also "locally admissible".

Theorem 3 shows that also complete semantics is fully decomposable[8]. Since the proof adopts $F_{\mathbf{CO}}$ as the local function and **CO** is complete-compatible,

by Proposition 5 it holds that $F_{\mathbf{CO}}$ is actually the canonical local function of complete semantics.

Theorem 3. *Complete semantics* **CO** *is fully decomposable and* $F_{\mathbf{CO}}$ *is its canonical local function.*

[8]Proposition 3 of [272] proves a weaker property of complete semantics, corresponding to bottom-up decomposability in the extension-based approach.

Stable semantics

Stable semantics inherits full decomposability from complete semantics: the reason is that the definition of stable labelling corresponds to that of complete labelling with the additional requirement that no argument is labelled undec, and this requirement holds at the level of the whole argumentation framework iff it holds in any of its subframeworks. The relevant local function can easily be identified by taking into account this requirement (again, the fact that such local function is the canonical one holds in virtue of Proposition 5).

Definition 27. Given an argumentation framework with input $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, $F_{\mathbf{ST}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}}) \triangleq \{Lab \in F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}}) \mid \forall A \in Ar, Lab(A) \neq \texttt{undec}\}$.

Theorem 4. *Stable semantics* **ST** *is fully decomposable and* $F_{\mathbf{ST}}$ *is its canonical local function.*

Example 4. Consider again the running example of Figure 2.1. Taking into account the results provided in Example 2 for the local function of complete semantics, it is easy to see that $F_{\mathbf{ST}}(AF{\downarrow_{\{A,B,C\}}}, \{D\}, \{(D, \texttt{out})\}, \{(D,A)\}) = \emptyset$, that $F_{\mathbf{ST}}(AF{\downarrow_{\{A,B,C\}}}, \{D\}, \{(D, \texttt{in})\}, \{(D,A)\}) = \{\{(A, \texttt{out}), (B, \texttt{in}), (C, \texttt{out})\}\}$, and for $AF{\downarrow_{\{D\}}}$ that $F_{\mathbf{ST}}(AF{\downarrow_{\{D\}}}, \{A\}, \{(A, \texttt{out})\}, \{(A,D)\}) = \{\{(D, \texttt{in})\}\}$, $F_{\mathbf{ST}}(AF{\downarrow_{\{D\}}}, \{A\}, \{(A, \texttt{in})\}, \{(A,D)\}) = \{\{(D, \texttt{out})\}\}$. Accordingly, there is just a pair of compatible local labellings, namely $\{(A, \texttt{out}), (B, \texttt{in}), (C, \texttt{out})\}$ and $\{(D, \texttt{in})\}$, giving rise to the unique stable labelling $\{(A, \texttt{out}), (B, \texttt{in}), (C, \texttt{out}), (D, \texttt{in})\}$.

Grounded and Preferred semantics

As in the previous cases, the canonical local functions of grounded and preferred semantics can be obtained by extending the definition of grounded and preferred labelling, respectively. Proposition 6 identifies these functions, also showing that the relevant definitions are well-founded, in particular, that there is always a unique minimal labelling in $F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ and that $F_{\mathbf{PR}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ is nonempty.

Proposition 6. The canonical local function of grounded and preferred semantics are defined as

- $F_{\mathbf{GR}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}}) \triangleq \{\mathbf{L}^*\}$, where $\mathbf{L}^*$ is the minimal (w.r.t. $\sqsubseteq$) labelling in $F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$
- $F_{\mathbf{PR}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}}) \triangleq \{\mathbf{L} \mid \mathbf{L}$ is a maximal (w.r.t. $\sqsubseteq$) labelling in $F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})\}$.

Differently from stable semantics, grounded semantics and preferred semantics do not inherit decomposability from complete semantics. The reason is that the definition of grounded/preferred labelling includes a minimization/maximization requirement, and satisfying this requirement in all of the subframeworks does not entail satisfying it at the level of the whole framework. To show this, consider the following counterexample.[9]

Example 5. We have shown in Example 2 that in the running example of Figure 2.1 the outcome of $F_{\mathbf{CO}}$ is a unique labelling in all cases, thus by definition it coincides with the outcome of $F_{\mathbf{GR}}$ and $F_{\mathbf{PR}}$. Given the compatibility constraint, exactly two global labellings arise from the combinations of the outcomes of $F_{\mathbf{CO}}$, namely $\{(A, \texttt{undec}), (B, \texttt{undec}), (C, \texttt{undec}), (D, \texttt{undec})\}$ and $\{(A, \texttt{out}), (B, \texttt{in}), (C, \texttt{out}), (D, \texttt{in})\}$. The former is the grounded labelling, the latter is the preferred labelling: it turns out that the combination of two "locally grounded" labellings gives rise not just to the "global" grounded labelling but also to the preferred labelling, and analogously that the combination of two "locally preferred" labellings gives rise not

[9] A counterexample to decomposability of grounded semantics is provided also in [272].

just to the "global" preferred labelling but also to the grounded one. This shows that grounded and preferred semantics are not bottom-up decomposable.

Now, a question arises as to whether satisfying the minimization/maximization requirement at the level of the whole argumentation framework entails that such requirement is satisfied at the local level, i.e. whether grounded and preferred semantics are top-down decomposable. This result turns out to be true and is achieved through some intermediate steps.

First, Lemma 1 shows that if a labelling produced by $F_{\mathbf{AD}}$ does not belong to $F_{\mathbf{CO}}$ then there is an `undec`-labelled argument which can be labelled `in` or `out` obtaining a labelling still in $F_{\mathbf{AD}}$.

Lemma 1. Given an argumentation framework with input $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, where $AF = (Ar, \rightarrow)$, let $\mathbf{L}$ be a labelling such that $\mathbf{L} \in F_{\mathbf{AD}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ and $\mathbf{L} \notin F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$. Then there is an argument $A \in Ar$ such that $\mathbf{L}(A) = \texttt{undec}$ and a labelling $\mathbf{L}^A \in F_{\mathbf{AD}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ such that $\mathbf{L}^A(A) \in \{\texttt{in}, \texttt{out}\}$ and $\forall B \in Ar : B \neq A$, $\mathbf{L}^A(B) = \mathbf{L}(B)$.

Lemma 2 shows that for every labelling produced by $F_{\mathbf{AD}}$ there is a more or equally committed labelling produced by $F_{\mathbf{CO}}$.

Lemma 2. Given an argumentation framework with input $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, for every labelling $\mathbf{L}_1 \in F_{\mathbf{AD}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ there exists a labelling $\mathbf{L}_2 \in F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ such that $\mathbf{L}_1 \sqsubseteq \mathbf{L}_2$.

Proposition 7 shows a sort of monotonicity property of $F_{\mathbf{CO}}$ with respect to the $\sqsubseteq$ relation.

Proposition 7. Given an argumentation framework with input $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, let $L_{\mathscr{I}}{}^1, L_{\mathscr{I}}{}^2 \in \mathfrak{L}_{\mathscr{I}}$ be two labellings of $\mathscr{I}$ such that $L_{\mathscr{I}}{}^1 \sqsubseteq L_{\mathscr{I}}{}^2$. Then it holds that

1. $\forall \mathbf{L}_1 \in F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}{}^1, R_{\mathscr{I}}), \exists \mathbf{L}_2 \in F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}{}^2, R_{\mathscr{I}})$ such that $\mathbf{L}_1 \sqsubseteq \mathbf{L}_2$; and
2. $\forall \mathbf{L}_2 \in F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}{}^2, R_{\mathscr{I}}), \exists \mathbf{L}_1 \in F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}{}^1, R_{\mathscr{I}})$ such that $\mathbf{L}_1 \sqsubseteq \mathbf{L}_2$.

Building on the above results (more specifically, using the first point for Theorem 6 and the second point for Theorem 5) we are now in a position to prove that grounded and preferred semantics are top-down decomposable.

Theorem 5. *Given an argumentation framework $AF = (Ar, \rightarrow)$, let $\mathbf{L}$ be the grounded labelling of AF. For any set $P \subseteq Ar$, $\mathbf{L}{\downarrow_P} \in F_{\mathbf{GR}}(AF{\downarrow_P}, P^{inp}, \mathbf{L}{\downarrow_{P^{inp}}}, P^R)$.*

Theorem 6. *Given an argumentation framework $AF = (Ar, \rightarrow)$, let $\mathbf{L}$ be a preferred labelling of AF. For any set $P \subseteq Ar$, $\mathbf{L}{\downarrow_P} \in F_{\mathbf{PR}}(AF{\downarrow_P}, P^{inp}, \mathbf{L}{\downarrow_{P^{inp}}}, P^R)$.*

While preferred and complete semantics fail to achieve bottom-up decomposability for arbitrary partitions, they turn out to be bottom-up decomposable (thus fully decomposable) w.r.t. $\mathscr{F}_{\cup\mathsf{SCC}}$. The result, proved in Theorem 7, is based on a preliminary lemma, which roughly states that if a semantics $\mathbf{S}$ is top-down decomposable then a kind of top-down decomposability relation holds for any labelling $\mathbf{L} \in F_{\mathbf{S}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ w.r.t. any set of arguments P in AF. More specifically, given such a labelling $\mathbf{L}$ and P, it is possible to refer to a "restricted" argumentation framework with input based on P, namely $(AF{\downarrow_P}, P^{\text{F-inp}}, (\mathbf{L} \cup L_{\mathscr{I}}){\downarrow_{P^{\text{F-inp}}}}, P_F^R)$, where intuitively $P^{\text{F-inp}}$, $(\mathbf{L} \cup L_{\mathscr{I}}){\downarrow_{P^{\text{F-inp}}}}$, and P_F^R are obtained by considering both $\mathscr{I}$ and AF (outside P). Then, the restriction of $\mathbf{L}$ to P is produced by $F_{\mathbf{S}}$ when applied to the restricted argumentation framework with input mentioned above.

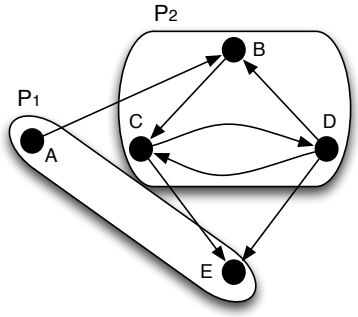


Figure 2.3: A partition belonging to $\mathscr{F}_{\cup \mathtt{SCC}}$ (Example 6).

Lemma 3. Let $\mathbf{S}$ be a complete-compatible semantics which is top-down decomposable, with the canonical local function $F_{\mathbf{S}}$. Given an argumentation framework with input $(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, consider a labelling $\mathbf{L} \in F_{\mathbf{S}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ and let $P \subseteq Ar$ be an arbitrary set of arguments of AF. Then, letting $P^{\text{F-inp}} \triangleq P^{\text{inp}} \cup \{A \in \mathscr{I} \mid \exists B \in P, (A,B) \in R_{\mathscr{I}}\}$ and $P_F^R \triangleq P^R \cup (R_{\mathscr{I}} \cap (\mathscr{I} \times P))$, it holds that $\mathbf{L}{\downarrow_P} \in F_{\mathbf{S}}(AF{\downarrow_P}, P^{\text{F-inp}}, (\mathbf{L} \cup L_{\mathscr{I}}){\downarrow_{P^{\text{F-inp}}}}, P_F^R)$.

Theorem 7. *Grounded and preferred semantics are decomposable w.r.t. $\mathscr{F}_{\cup \mathtt{SCC}}$.*

Example 6. Consider $AF = (\{A,B,C,D,E\}, \{(A,B),(B,C),(C,D),(D,C),(D,B),(C,E),(D,E)\})$ and the partition $\{P_1,P_2\} \in \mathscr{F}_{\cup \mathtt{SCC}}(AF)$ where $P_1 = \{A,E\}$ and $P_2 = \{B,C,D\}$ (see Figure 2.3). It holds that ${P_1}^{\text{inp}} = \{C,D\}$, ${P_1}^R = \{(C,E),(D,E)\}$, ${P_2}^{\text{inp}} = \{A\}$, ${P_2}^R = \{(A,B)\}$. Note that the partition is not "acyclic", in that P_1 attacks P_2 and P_2 attacks P_1. We show that both in the case of grounded semantics and of preferred semantics the union of compatible local labellings gives rise to the grounded labelling or a preferred labelling, respectively. First, note that any labelling returned by $F_{\mathbf{GR}}$ and $F_{\mathbf{PR}}$ applied to $AF{\downarrow_{P_1}}$ prescribes that A is labelled $\mathtt{in}$, therefore it suffices to consider the labelling $\{(A,\mathtt{in})\}$ for the unique input argument of P_2. As to grounded semantics, it turns out that $F_{\mathbf{GR}}(AF{\downarrow_{P_2}}, \{A\}, \{(A,\mathtt{in})\}, \{(A,B)\}) = \{\{(B,\mathtt{out}),(C,\mathtt{undec}),(D,\mathtt{undec})\}\}$, while $F_{\mathbf{GR}}(AF{\downarrow_{P_1}}, \{C,D\}, \{(C,\mathtt{undec}),(D,\mathtt{undec})\}, \{(C,E),(D,E)\}) = \{\{(A,\mathtt{in}),(E,\mathtt{undec})\}\}$. We have a unique pair of compatible local labellings which give rise to the global labelling $\{(A,\mathtt{in}),(B,\mathtt{out}),(C,\mathtt{undec}),(D,\mathtt{undec}),(E,\mathtt{undec})\}$, i.e. the grounded labelling of AF. As to preferred semantics, $F_{\mathbf{PR}}(AF{\downarrow_{P_2}}, \{A\}, \{(A,\mathtt{in})\}, \{(A,B)\})$ returns two labellings, i.e. $\{(B,\mathtt{out}),(C,\mathtt{in}),(D,\mathtt{out})\}$ and $\{(B,\mathtt{out}),(C,\mathtt{out}),(D,\mathtt{in})\}$, while $F_{\mathbf{PR}}(AF{\downarrow_{P_1}}, \{C,D\}, \{(C,\mathtt{in}),(D,\mathtt{out})\}, \{(C,E),(D,E)\}) = F_{\mathbf{PR}}(AF{\downarrow_{P_1}}, \{C,D\}, \{(C,\mathtt{out}),(D,\mathtt{in})\}, \{(C,E),(D,E)\}) = \{\{(A,\mathtt{in}),(E,\mathtt{out})\}\}$. Accordingly, the union of compatible local labellings gives rise to $\{(A,\mathtt{in}),(B,\mathtt{out}),(C,\mathtt{in}),(D,\mathtt{out}),(E,\mathtt{out})\}$ and $\{(A,\mathtt{in}),(B,\mathtt{out}),(C,\mathtt{out}),(D,\mathtt{in}),(E,\mathtt{out})\}$, i.e. the preferred labellings of AF.

Ideal semantics

Similarly to the cases analyzed in the previous sections, the canonical local function of ideal semantics corresponds to an extension of the definition of ideal labelling. The following proposition identifies the rel-

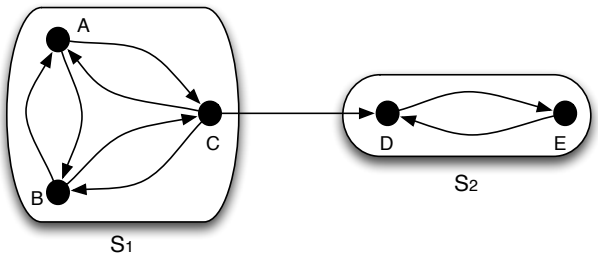


Figure 2.4: Ideal semantics is neither top-down nor bottom-up decomposable w.r.t. $\mathscr{F}_{\mathsf{SCC}}$ (Example 7).

evant definition, also showing that it is well founded (in particular, that $F_{\mathbf{ID}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ always returns a unique labelling).

Proposition 8. The canonical local function of ideal semantics is defined as $F_{\mathbf{ID}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}}) \triangleq \{\mathbf{L}^*\}$, where $\mathbf{L}^*$ is the maximal (w.r.t. $\sqsubseteq$) labelling in $F_{\mathbf{CO}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ such that for each $\mathbf{L}_P \in F_{\mathbf{PR}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ it holds that $\mathbf{L}^* \sqsubseteq \mathbf{L}_P$.

Ideal semantics has some common features both with preferred and with grounded semantics: on the one hand, its definition is based on the preferred labellings, on the other hand it yields a unique labelling, like the grounded semantics. As a matter of fact, a formal skepticism comparison between semantics shows that ideal semantics lies between grounded and preferred semantics [16]. Ideal semantics does not inherit any decomposability property from them: the following example shows that ideal semantics is neither top-down nor bottom-up decomposable even w.r.t. $\mathscr{F}_{\mathsf{SCC}}$.

Example 7. $AF = (\{A,B,C,D,E\}, \{(A,B),(B,A),(A,C),(C,A),(B,C),(C,B),(C,D),(D,E),(E,D)\})$ has the unique partition $\{S_1, S_2\} \in \mathscr{F}_{\mathsf{SCC}}(AF)$, where $S_1 = \{A,B,C\}$ and $S_2 = \{D,E\}$ are the strongly connected components of AF (see Figure 2.4). There are 5 preferred labellings of AF and there is no argument which is labelled `in` in all of them, thus the ideal labelling $\mathbf{L}^*$ leaves all of the arguments undecided. To show that ideal semantics is not top-down decomposable w.r.t. $\mathscr{F}_{\mathsf{SCC}}$, it is sufficient to note that $\mathbf{L}^*{\downarrow_{S_2}} = \{(D, \mathtt{undec}), (E, \mathtt{undec})\}$, while it turns out that $F_{\mathbf{ID}}(AF{\downarrow_{S_2}}, \{C\}, \{(C, \mathtt{undec})\}, \{(C,D)\}) = \{\{(D, \mathtt{out}), (E, \mathtt{in})\}\}$.

To show that ideal semantics is not bottom-up decomposable w.r.t. $\mathscr{F}_{\mathsf{SCC}}$, consider first the application of $F_{\mathbf{ID}}$ to $AF{\downarrow_{S_1}}$: it is easy to see that $F_{\mathbf{ID}}(AF{\downarrow_{S_1}}, \emptyset, \emptyset, \emptyset) = \{(A, \mathtt{undec}), (B, \mathtt{undec}), (C, \mathtt{undec})\}$, since $AF{\downarrow_{S_1}}$ admits the three preferred labellings where one of the three arguments $\{A,B,C\}$ is `in` and the others are `out`. Moreover, we already know that $F_{\mathbf{ID}}(AF{\downarrow_{S_2}}, \{C\}, \{(C, \mathtt{undec})\}, \{(C,D)\}) = \{\{(D, \mathtt{out}), (E, \mathtt{in})\}\}$, thus the labellings $\{(A, \mathtt{undec}), (B, \mathtt{undec}), (C, \mathtt{undec})\}$ and $\{\{(D, \mathtt{out}), (E, \mathtt{in})\}\}$ are compatible. However, the union of these two labellings does not coincide with the ideal labelling $\mathbf{L}^*$.

The previous example contradicts[10] a result presented in [201], according to which ideal semantics is decomposable w.r.t. partitions including two elements one of which is unattacked (i.e. does not receive attacks from outside, S_1 in Figure 2.4). The reason why ideal semantics is not decomposable is that, considering a strongly connected component P, the restriction of the ideal labelling to the input arguments of P does not always carry enough information to compute the restriction of the ideal labelling to P. In the

[10] A detailed discussion of this matter is given in [18].

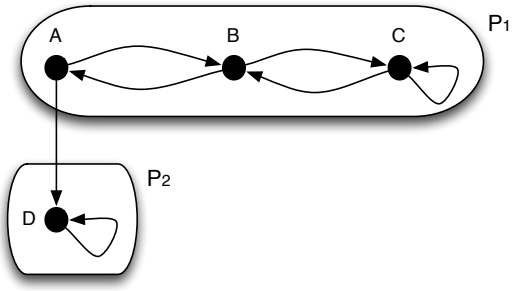


Figure 2.5: Semi-stable semantics is not top-down decomposable w.r.t. $\mathscr{F}_{\mathtt{SCC}}$ (Example 8).

previous example, argument C is labelled `undec` by the ideal labelling while it is labelled `in` or `out` by the preferred labellings, i.e. those which actually determine the ideal labelling according to Definition 10.

Semi-stable semantics

The definition of semi-stable semantics somewhat resembles that of preferred semantics, in that semi-stable labellings correspond to those preferred labellings which satisfy the additional requirement of minimizing the set of arguments labelled `undec`. The following proposition shows that the canonical local function is defined accordingly.

Proposition 9. The canonical local function of semi-stable semantics is defined as $F_{\mathbf{SST}}(AF,\mathscr{I},L_{\mathscr{I}},R_{\mathscr{I}}) \triangleq \{\mathbf{L} \mid \mathbf{L} \in F_{\mathbf{CO}}(AF,\mathscr{I},L_{\mathscr{I}},R_{\mathscr{I}})$ such that $\mathtt{undec}(\mathbf{L})$ is minimal w.r.t. set inclusion$\}$.

Differently from all semantics considered above, semi-stable semantics is not directional [15], i.e. given an unattacked set of arguments S the labellings computed in $AF{\downarrow_S}$ do not correspond to the restrictions of the labellings of AF in S. As shown in the following two examples, this behavior prevents the satisfaction of top-down and bottom-up decomposability even w.r.t. $\mathscr{F}_{\mathtt{SCC}}$.

Example 8. To show that semi-stable semantics is not top-down decomposable w.r.t. $\mathscr{F}_{\mathtt{SCC}}$, consider $AF = (\{A,B,C,D\},\{(A,B),(B,A),(B,C),(C,B),(C,C),(A,D),(D,D)\})$, where $\mathrm{SCCS}_{AF} = \{P_1,P_2\}$ with $P_1 = \{A,B,C\}$ and $P_2 = \{D\}$ (see Figure 2.5). There are two semi-stable labellings in AF, namely $\mathbf{L}_1 = \{(A,\mathtt{in}),(B,\mathtt{out}),(C,\mathtt{undec}),(D,\mathtt{out})\}$ and $\mathbf{L}_2 = \{(A,\mathtt{out}),(B,\mathtt{in}),(C,\mathtt{out}),(D,\mathtt{undec})\}$. Consider then the partition $\{P_1,P_2\} \in \mathscr{F}_{\mathtt{SCC}}(AF)$ where P_1 is unattacked. Note in particular that $\mathbf{L}_1{\downarrow_{P_1}} = \{(A,\mathtt{in}),(B,\mathtt{out}),(C,\mathtt{undec})\}$, which however does not belong to $F_{\mathbf{SST}}(AF{\downarrow_{P_1}},\emptyset,\emptyset,\emptyset)$, since the only semi-stable labelling in $AF{\downarrow_{P_1}}$ is $\{(A,\mathtt{out}),(B,\mathtt{in}),(C,\mathtt{out})\}$.

Example 9. To show that semi-stable semantics is not bottom-up decomposable w.r.t. $\mathscr{F}_{\mathtt{SCC}}$, consider the argumentation framework $AF = (\{A,B,C\},\{(A,B),(B,A),(B,C),(C,C)\})$ and the partition $\{P_1,P_2\} \in \mathscr{F}_{\mathtt{SCC}}(AF)$ with $P_1 = \{A,B\}$ and $P_2 = \{C\}$ (see Figure 2.6). It is easy to see that $\{(A,\mathtt{in}),(B,\mathtt{out})\} \in F_{\mathbf{SST}}(AF{\downarrow_{P_1}},\emptyset,\emptyset,\emptyset)$, and that $F_{\mathbf{SST}}(AF{\downarrow_{P_2}},\{B\},\{(B,\mathtt{out})\},\{(B,C)\}) = \{\{(C,\mathtt{undec})\}\}$. Now, the union of these compatible labellings, i.e. $\{(A,\mathtt{in}),(B,\mathtt{out}),(C,\mathtt{undec})\}$, is not a semi-stable labelling of AF, since the unique semi-stable labelling of AF is $\{(A,\mathtt{out}),(B,\mathtt{in}),(C,\mathtt{out})\}$.

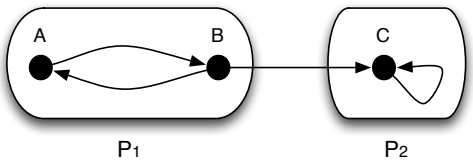


Figure 2.6: Semi-stable semantics is not bottom-up decomposable w.r.t. $\mathscr{F}_{\mathtt{SCC}}$ (Example 9).

2.6 Effect-dictated semantics

This short section introduces the simple concept of *effect-dictated* semantics, which is crucial for the analysis to be carried out in the next section. For every semantics **S** analyzed in Section 2.5, it can be noted that $F_{\mathbf{S}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ may return the same result given different $\mathscr{I}$, $L_{\mathscr{I}}$ and $R_{\mathscr{I}}$. For instance, if an argument A of AF is attacked by an argument of $\mathscr{I}$ which is labelled `in`, then $F_{\mathbf{S}}$ returns the same set of labellings independently of the presence and the number of additional attackers of A in $\mathscr{I}$. The *effect* of $(\mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ on the arguments $Args$ of AF can be modelled as the labelling that would be induced on $Args$ by neglecting the attacks inside AF. For instance, if an argument A of AF is only attacked through $R_{\mathscr{I}}$ by `out`-labelled arguments according to $L_{\mathscr{I}}$, then A would be `in` in the case that it does not receive other attacks inside AF.[11] The following definition formalizes this intuition.

Definition 28. Given a set of arguments $\mathscr{I}$, a labelling $L_{\mathscr{I}} \in \mathfrak{L}_{\mathscr{I}}$, a set of arguments $Args$ such that $\mathscr{I} \cap Args = \emptyset$ and a relation $R_{INP} \subseteq \mathscr{I} \times Args$, the effect of $(\mathscr{I}, L_{\mathscr{I}}, R_{INP})$ on $Args$, denoted as $\mathtt{eff}_{Args}(\mathscr{I}, L_{\mathscr{I}}, R_{INP})$, is defined as

$$\{(A, \mathtt{out}) \mid A \in Args, \exists B \in \mathscr{I} : (B,A) \in R_{INP} \wedge L_{\mathscr{I}}(B) = \mathtt{in}\} \cup$$
$$\{(A, \mathtt{undec}) \mid A \in Args, \exists B \in \mathscr{I} : (B,A) \in R_{INP} \wedge L_{\mathscr{I}}(B) = \mathtt{undec}, \nexists C \in \mathscr{I} : (C,A) \in R_{INP} \wedge L_{\mathscr{I}}(C) = \mathtt{in}\} \cup$$
$$\{(A, \mathtt{in}) \mid A \in Args, \nexists B \in \mathscr{I} : (B,A) \in R_{INP} \wedge L_{\mathscr{I}}(B) \in \{\mathtt{in}, \mathtt{undec}\}\}$$

By definition, $\mathtt{eff}_{Args}(\mathscr{I}, L_{\mathscr{I}}, R_{INP})$ only depends on the labelling of the arguments in $\mathscr{I}$ that attack $Args$ through R_{INP}. Moreover each argument in $Args$ not receiving attacks from $\mathscr{I}$ is labelled `in` according to $\mathtt{eff}_{Args}(\mathscr{I}, L_{\mathscr{I}}, R_{INP})$. Thus, in the particular case where $\mathscr{I} = \emptyset$ (thus also $L_{\mathscr{I}}$ and R_{INP} are empty), it turns out that $\mathtt{eff}_{Args}(\emptyset, \emptyset, \emptyset) = \{(A, \mathtt{in}) \mid A \in Args\}$.

The following lemma proves a monotonic relation between labellings and effects.

Lemma 4. Given a set of arguments $\mathscr{I}$, two labellings $L^1_{\mathscr{I}}, L^2_{\mathscr{I}} \in \mathfrak{L}_{\mathscr{I}}$, a set of arguments $Args$ such that $\mathscr{I} \cap Args = \emptyset$ and a relation $R_{INP} \subseteq \mathscr{I} \times Args$, if $L^1_{\mathscr{I}} \sqsubseteq L^2_{\mathscr{I}}$ then $\mathtt{eff}_{Args}(\mathscr{I}, L^1_{\mathscr{I}}, R_{INP}) \sqsubseteq \mathtt{eff}_{Args}(\mathscr{I}, L^2_{\mathscr{I}}, R_{INP})$.

A semantics **S** is said to be *effect-dictated* if, given $AF = (Ar, \rightarrow)$, $F_{\mathbf{S}}(AF, \mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$ only depends on $\mathtt{eff}_{Ar}(\mathscr{I}, L_{\mathscr{I}}, R_{\mathscr{I}})$, rather than on the whole labelling $L_{\mathscr{I}}$ and the specific relation $R_{\mathscr{I}}$.

Definition 29. A semantics **S** is effect-dictated if $(\mathtt{eff}_{Ar}(\mathscr{I}_1, L_{\mathscr{I}_1}, R_{\mathscr{I}_1}) = \mathtt{eff}_{Ar}(\mathscr{I}_2, L_{\mathscr{I}_2}, R_{\mathscr{I}_2})) \Rightarrow F_{\mathbf{S}}(AF, \mathscr{I}_1, L_{\mathscr{I}_1}, R_{\mathscr{I}_1}) = F_{\mathbf{S}}(AF, \mathscr{I}_2, L_{\mathscr{I}_2}, R_{\mathscr{I}_2})$ for every AF, $\mathscr{I}_1$, $\mathscr{I}_2$, $L_{\mathscr{I}_1}$, $L_{\mathscr{I}_2}$, $R_{\mathscr{I}_1}$ and $R_{\mathscr{I}_2}$, where $AF = (Ar, \rightarrow)$ is an argumentation framework, $\mathscr{I}_1$ and $\mathscr{I}_2$ are two sets of arguments such that $\mathscr{I}_1 \cap Ar = \emptyset$ and

[11] The effect is a bit different for admissible semantics, but this does not affect its technical treatment, as well as the subsequent results.

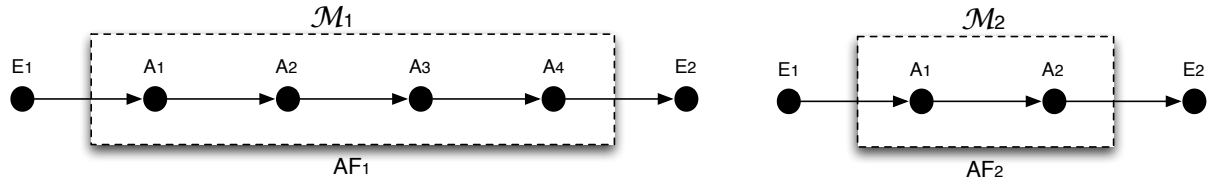


Figure 2.7: Summarizing a chain of arguments (Example 10).

$\mathscr{I}_2 \cap Ar = \emptyset$, $L_{\mathscr{I}_1} \in \mathfrak{L}_{\mathscr{I}_1}$ and $L_{\mathscr{I}_2} \in \mathfrak{L}_{\mathscr{I}_2}$ two labellings of $\mathscr{I}_1$ and $\mathscr{I}_2$ respectively, and $R_{\mathscr{I}_1} \subseteq \mathscr{I}_1 \times Ar$ and $R_{\mathscr{I}_2} \subseteq \mathscr{I}_2 \times Ar$ two relations.

All the semantics considered in this chapter are effect-dictated as shown by the following proposition.

Proposition 10. Every semantics $\mathbf{S} \in \{\mathbf{AD}, \mathbf{CO}, \mathbf{ST}, \mathbf{GR}, \mathbf{PR}, \mathbf{ID}, \mathbf{SST}\}$ is effect-dictated.

2.7 Argumentation Multipoles and their interchangeability

In this section, we introduce argumentation multipoles, that are conceived as modular components equipped with a well-defined interface to connect with each other and may play the role of "partial" frameworks in the context of a global one. This yields the possibility of replacing a component with another one which is equivalent as far as the Input/Output behavior is concerned.

The notion of Argumentation Multipole

The first step to provide a systematic treatment of argumentation multipoles is to identify a definition to capture their structure in the most general way. To this aim, we consider a number of examples, starting from a common component, i.e. a chain of arguments.

Example 10. Consider the argumentation frameworks AF_1 and AF_2 shown in Figure 2.7. AF_2 can be obtained from AF_1 by "summarizing" the component $\mathscr{M}_1$, including the arguments A_1, A_2, A_3, A_4, with the component $\mathscr{M}_2$, including the arguments A_1 and A_2: according to any complete-compatible semantics considered in this chapter, the labellings restricted to E_1 and E_2, i.e. the arguments common to AF_1 and AF_2, are the same in the two frameworks, i.e. E_1 is labelled `in` and E_2 is labelled `out`. More generally, consider a finite sequence of n arguments $A_1, \ldots, A_n$ such that each argument attacks the subsequent one, i.e. A_i attacks A_{i+1} with $1 \leq i < n$ and suppose that only A_1 can receive further attacks from other arguments and only A_n can attack other arguments. Then it is intuitive to see that the "black-box behavior" of a sequence of arguments of this kind, whose external "terminals" are A_1 and A_n, only depends on whether n is even or odd. In fact, the behavior of any even-length sequence is the same as in the case $n = 2$ (if A_1 is `in` then A_n is `out`, if A_1 is `out` then A_n is `in`, if A_1 is `undec` then A_n is `undec`),

while for any odd-length sequence the behavior is the same as the one of A_1 alone (with n odd, A_n gets necessarily the same label as A_1).

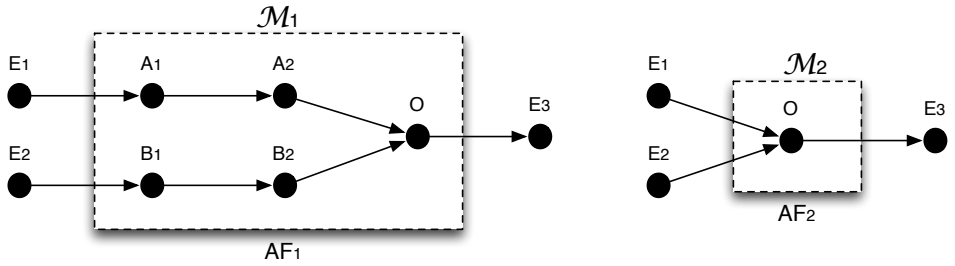


Figure 2.8: Summarizing two chains of arguments attacking an argument O (Example 11).

On the basis of the previous example, a modular component may tentatively be defined as an argumentation framework[12] where the "input terminals" and the "output terminals" are explicitly identified (e.g. $AF_1{\downarrow}_{\{A_1,\ldots,A_4\}}$ in the example, where A_1 is the unique input terminal and A_4 in the unique output terminal). Two components can be interchanged only if they have the same input and output terminals, and this interchange does not modify the attacks relating these terminals with the unchanged arguments (E_1 and E_2 in the example). However, the following two examples show that this approach is too restrictive, since there are cases where it is useful to modify both the set of input and output terminals as well as the relevant attack relation.

Example 11. Consider the argumentation frameworks AF_1 and AF_2 shown in Figure 2.8. AF_2 can be obtained from AF_1 by summarizing the component $\mathscr{M}_1$, including the arguments A_1, A_2, B_1, B_2, O, with the component $\mathscr{M}_2$ including the argument O only: according to all complete-compatible semantics considered in this chapter the arguments E_1, E_2 and E_3 are labelled `in` both in AF_1 and AF_2. More generally, the black-box behavior of $\mathscr{M}_1$ is the same as the one of $\mathscr{M}_2$, since in $\mathscr{M}_1$ A_2 gets the same label as E_1 and B_2 gets the same label as E_2, thus the label of O is the same as in $\mathscr{M}_2$. As a consequence, one may expect that $\mathscr{M}_1$ can be interchanged with $\mathscr{M}_2$ also in more articulated examples. Note that while $\mathscr{M}_1$ has two input terminals, $\mathscr{M}_2$ has only one input terminal coinciding with the unique output one.

Example 12. Consider the argumentation frameworks AF_1 and AF_2 shown in Figure 2.9 and assume preferred semantics is adopted. AF_2 can be obtained from AF_1 by summarizing the component $\mathscr{M}_1$, including the arguments A_1, A_2 and O, with the component $\mathscr{M}_2$ including the arguments I and O: both in AF_1 and AF_2 the argument E_1 is labelled `in`, E_2 is labelled `out` and E_3 is labelled `in`. More generally, under preferred semantics the black-box behavior of $\mathscr{M}_1$ is the same as the one of $\mathscr{M}_2$: if E_2 is `in` then O is `in`, if E_2 is `out` then O is `out` (in particular $\mathscr{M}_1$ admits a labelling where A_1 is `in`, A_2 is `out` and O is `out`, and a labelling where A_1 is `out`, A_2 is `in` and O is `out`), if E_2 is `undec` then O is `undec`. As a consequence, one may expect that $\mathscr{M}_1$ can be interchanged with $\mathscr{M}_2$ also in more articulated examples. Note that while $\mathscr{M}_1$ receives two attacks from E_2 in AF_1, $\mathscr{M}_2$ receives one attack only in AF_2.

The previous examples show that the definition of a modular component should include the *input attack relation* R_{INP}, consisting of the attacks from the arguments that are not part of the component to the arguments that belong to the component itself: this way, the definition leaves room for replacements of modular

[12] This approach has been followed in our chapter [23], leading to the notion of *Input/Output Argumentation Framework*.

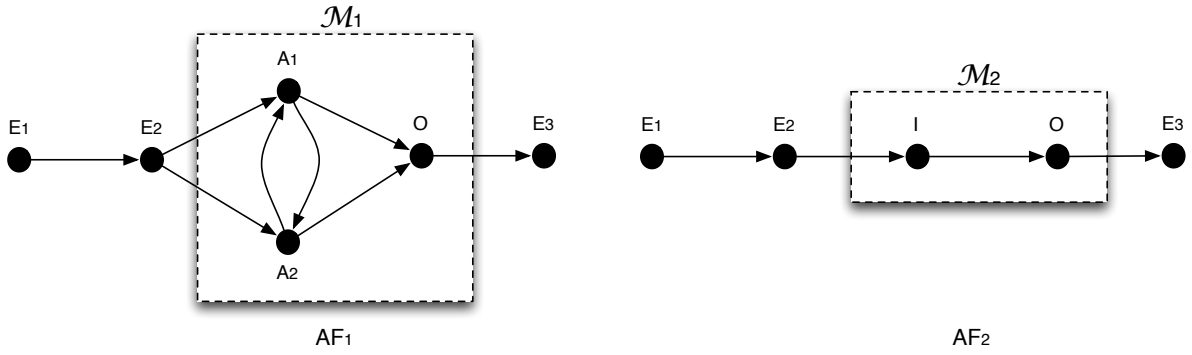


Figure 2.9: Summarizing two contradicting arguments attacking an argument O (Example 12).

components that lead to changes in the input attack relation, as in the previous example. A similar reasoning concerns the *output attack relation* R_{OUTP}, including the attacks from a modular component towards the outside arguments. In any case, there is no need to explicitly model the input and output terminals, since they can easily be derived from the input and output attack relations. Inspired by the digital logic field, we call the resulting structure an *Argumentation Multipole*. In order to express R_{INP} and R_{OUTP}, without loss of generality we define an Argumentation Multipole *w.r.t. a set* E, i.e. w.r.t. the set of arguments that are not part of the multipole and thus remain unchanged if the multipole is replaced.

Definition 30. An Argumentation Multipole (or, briefly, multipole) $\mathscr{M}$ w.r.t. a set E is a tuple (AF, R_{INP}, R_{OUTP}), where letting $AF = (Ar, \rightarrow)$ it holds that $Ar \cap E = \emptyset$, $R_{INP} \subseteq E \times Ar$, and $R_{OUTP} \subseteq Ar \times E$. Extending the notation introduced in Definition 12, we denote as $\mathscr{M}^{\text{inp}}$ the set $\{A \in E \mid \exists B \in Ar, (A,B) \in R_{INP}\}$, i.e. including the arguments of E which attack Ar through R_{INP}. Moreover, we denote as $\mathscr{M}^{\text{outp}}$ the set $\{A \in Ar \mid \exists B \in E, (A,B) \in R_{OUTP}\}$, i.e. including the arguments of AF attacking E through R_{OUTP}.

Figure 2.10 provides a graphical representation of the definition. For instance, in Example 10 $\mathscr{M}_1 = (AF_1{\downarrow}_{\{A_1,A_2,A_3,A_4\}}, \{(E_1,A_1)\}, \{(A_4,E_2)\})$ and $\mathscr{M}_2 = (AF_2{\downarrow}_{\{A_1,A_2\}}, \{(E_1,A_1)\}, \{(A_2,E_2)\})$, in Example 11 it holds that $\mathscr{M}_1 = (AF_1{\downarrow}_{\{A_1,A_2,B_1,B_2,O\}}, \{(E_1,A_1),(E_2,B_1)\}, \{(O,E_3)\})$ and $\mathscr{M}_2 = (AF_2{\downarrow}_{\{O\}}, \{(E_1,O),(E_2,O)\}, \{(O,E_3)\})$, in Example 12 $\mathscr{M}_1 = (AF_1{\downarrow}_{\{A_1,A_2,O\}}, \{(E_2,A_1), (E_2,A_2)\}, \{(O,E_3)\})$ and $\mathscr{M}_2 = (AF_2{\downarrow}_{\{I,O\}}, \{(E_2,I)\}, \{(O,E_3)\})$.

A particular multipole which is useful to consider in some practical examples is the *empty multipole* $\mathscr{M}_\emptyset \triangleq (AF_\emptyset, \emptyset, \emptyset)$, i.e. including the empty argumentation framework $AF_\emptyset$. It is easy to see that $\mathscr{M}_\emptyset{}^{\text{inp}} = \mathscr{M}_\emptyset{}^{\text{outp}} = \emptyset$.

Input/Output equivalence of Argumentation Multipoles

After having introduced the definition of argumentation multipole, the next step is to formally characterize the relevant "black-box behavior": this way, the Input/Output equivalence relation between multipoles can be identified as the one relating the multipoles having the same behavior.

When a multipole w.r.t. a set E is "connected to the external world" it "receives" some input from outside through the relation R_{INP} and "produces" an output which is induced by the labellings of the multipole and transferred to the set E through the relation R_{OUTP}. Technically speaking, the labellings and thus

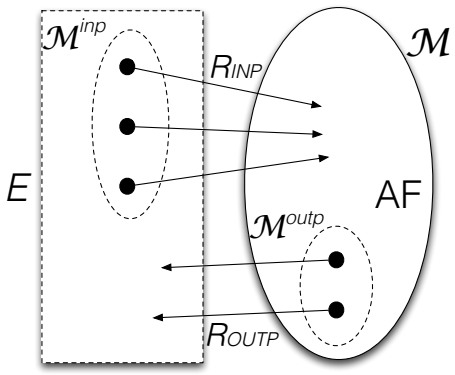


Figure 2.10: A graphical representation of the notion of argumentation multipole.

the relation between input and output are determined by a (semantics specific) local function, thus the equivalence relation between argumentation multipoles depends on the considered semantics **S**, and is called **S**-equivalence to reflect this dependency. For instance, in Example 12 $\mathscr{M}_1$ and $\mathscr{M}_2$ are **PR**-equivalent (i.e. equivalent under preferred semantics), while they are not **GR**-equivalent, since under grounded semantics if E_2 is labelled `out` then O in $\mathscr{M}_1$ is labelled `undec`, while O in $\mathscr{M}_2$ is labelled `out`. Intuitively, $\mathscr{M}_1$ is **GR**-equivalent e.g. to a multipole $\mathscr{M}_2'$ obtained from $\mathscr{M}_2$ by adding a self-attack from I to I itself.

According to the above examples, two argumentation multipoles w.r.t. the same set E may be tentatively defined as **S**-equivalent if for any possible input, i.e. any labelling of E, $F_{\mathbf{S}}$ produces the same labellings of the output terminals in the two argumentation multipoles. For instance, in Example 12 under preferred semantics O is `in` for any labelling where E_2 is `in`, it is `out` for any labelling where E_2 is `out` and it is `undec` for any labelling where E_2 is `undec`. However, this approach works only in case the two multipoles have the same output terminals. Moreover, as the following example shows, the way E is affected by the labellings of an argumentation multipole (AF, R_{INP}, R_{OUTP}) also depends on the attack relation R_{OUTP}.

Example 13. Consider the argumentation frameworks AF_1 and AF_2 shown in Figure 2.11 and the application of preferred semantics. The multipole $\mathscr{M}_1 = (AF_1{\downarrow}_{\{O_1,O_2\}}, \emptyset, \{(O_1,E),(O_2,E)\})$ w.r.t. $\{E\}$ in AF_1 affects the argument E by means of the two arguments O_1 and O_2, while $\mathscr{M}_2 = (AF_2{\downarrow}_{\{O\}}, \emptyset, \{(O,E)\})$ in AF_2 affects E by means of the argument O. Intuitively, under preferred semantics $\mathscr{M}_1$ and $\mathscr{M}_2$ are equivalent: in $\mathscr{M}_1$ there are two preferred labellings, i.e. $\{(O_1,\texttt{in}),(O_2,\texttt{out})\}$ and $\{(O_1,\texttt{out}),(O_2,\texttt{in})\}$, thus in any case an argument labelled `in` attacks E making it `out`, and similarly $\mathscr{M}_2$ interacts with E making it `out`, since $\mathscr{M}_2$ admits the unique labelling $\{(O,\texttt{in})\}$.

We can formalize these intuitions by extending the notion of effect to multipoles (see Definition 28). Let us consider a semantics **S**. Given a multipole $\mathscr{M}$ w.r.t. a set E, for any "input" labelling $\mathbf{L}_E \in \mathfrak{L}_E$ the local function $F_{\mathbf{S}}$ prescribes a set of labellings for $\mathscr{M}$. Each of these labellings has its own effect on E, therefore the global effect of the multipole receiving an input $\mathbf{L}_E$ is a set of labellings of E whose members are all the single effects.

Definition 31. Let $\mathscr{M} = (AF, R_{INP}, R_{OUTP})$ a multipole w.r.t. a set E and **S** an argumentation semantics. Given a labelling $\mathbf{L}_E \in \mathfrak{L}_E$, the **S**-effect of $(\mathscr{M}, \mathbf{L}_E)$ on E, denoted as $\mathbf{S}\text{-}\mathtt{eff}_E(\mathscr{M}, \mathbf{L}_E)$, is defined as $\{\mathtt{eff}_E(\mathscr{M}^{\text{outp}}, \mathbf{L}{\downarrow}_{\mathscr{M}^{\text{outp}}}, R_{OUTP}) \mid \mathbf{L} \in F_{\mathbf{S}}(AF, \mathscr{M}^{\text{inp}}, \mathbf{L}_E{\downarrow}_{\mathscr{M}^{\text{inp}}}, R_{INP})\}$.

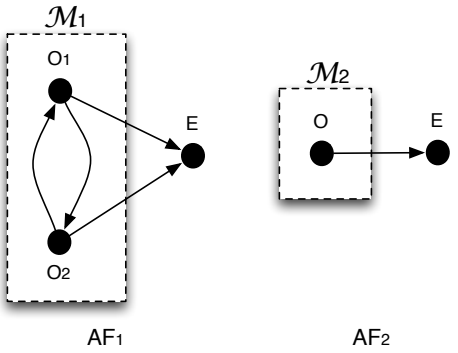


Figure 2.11: Summarizing two contradicting arguments (Example 13).

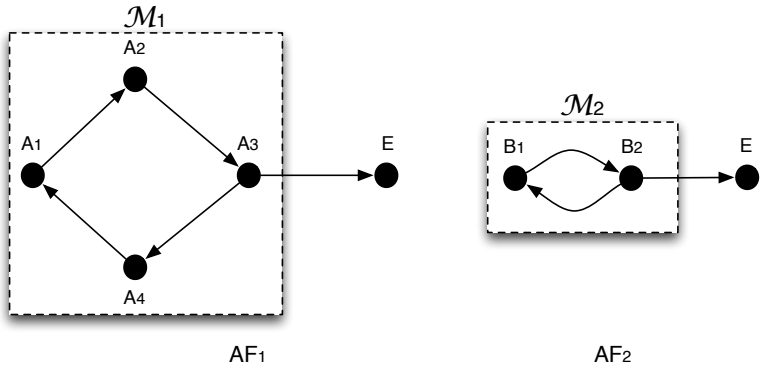


Figure 2.12: Summarizing a 4-length cycle of arguments (Example 14).

Note that if $F_{\mathbf{S}}(AF, \mathscr{M}^{\text{inp}}, \mathbf{L}_E{\downarrow}_{\mathscr{M}^{\text{inp}}}, R_{INP}) = \emptyset$, i.e. the local function prescribes no labelling, then $\mathbf{S}\text{-}\mathtt{eff}_E(\mathscr{M}, \mathbf{L}_E) = \emptyset$.

Example 14. Consider the argumentation frameworks AF_1 and AF_2 shown in Figure 2.12, and the multipoles $\mathscr{M}_1 = (AF_1{\downarrow}_{\{A_1,A_2,A_3,A_4\}}, \emptyset, \{(A_3, E)\})$ w.r.t. $\{E\}$ and $\mathscr{M}_2 = (AF_2{\downarrow}_{\{B_1,B_2\}}, \emptyset, \{(B_2, E)\})$ w.r.t. $\{E\}$. $\mathscr{M}_1$ has two preferred labellings, one where A_3 is `in` and another where A_3 is `out`, hence $\mathbf{PR}\text{-}\mathtt{eff}_{\{E\}}(\mathscr{M}_1, \emptyset) = \{\{(E_1, \mathtt{in})\}, \{(E_1, \mathtt{out})\}\}$. Similarly, $\mathscr{M}_2$ has two preferred labellings, one where B_2 is `in` and another where B_2 is `out`, leading to $\mathbf{PR}\text{-}\mathtt{eff}_{\{E\}}(\mathscr{M}_1, \emptyset) = \mathbf{PR}\text{-}\mathtt{eff}_{\{E\}}(\mathscr{M}_2, \emptyset)$.

It is worth considering the effect of the empty multipole $\mathscr{M}_\emptyset$. Intuitively, $\mathscr{M}_\emptyset$ should have no effect on the arguments of E, i.e. all of them should be assigned the label `in` according to the effect itself. Technically, this is guaranteed if the semantics is defined in such a way as to prescribe the unique possible labelling $\emptyset$ to the empty argumentation framework $AF_\emptyset$, as it happens for any semantics considered in this chapter. Intuitively, if this were not the case the empty multipole would prevent the identification of any labelling for the whole argumentation framework, yielding to a pathological behavior. Accordingly, the condition

$\mathbf{L_S}(AF_\emptyset) = \{\emptyset\}$ is required in all the following propositions and theorems[13] referring to a generic semantics **S**.

Proposition 11. Consider a semantics **S** such that $\mathbf{L_S}(AF_\emptyset) = \{\emptyset\}$. Given a set of arguments E and a labelling $\mathbf{L}_E \in \mathfrak{L}_E$, it holds that $\mathbf{S}\text{-}\mathtt{eff}_E(\mathscr{M}_\emptyset, \mathbf{L}_E) = \{\{(A, \mathtt{in}) \mid A \in E\}\}$.

Two multipoles $\mathscr{M}_1$ and $\mathscr{M}_2$ w.r.t. E can be considered **S**-equivalent if, for any possible labelling $\mathbf{L}_E \in \mathfrak{L}_E$, $\mathbf{S}\text{-}\mathtt{eff}_E(\mathscr{M}_1, \mathbf{L}_E) = \mathbf{S}\text{-}\mathtt{eff}_E(\mathscr{M}_2, \mathbf{L}_E)$. For reasons that will be clear later, it is also useful to identify multipoles that have the same effect only for a subset of input labellings: in order to capture this possibility, we define equivalence under a set of labellings of E.

Definition 32. Two multipoles $\mathscr{M}_1$ and $\mathscr{M}_2$ w.r.t. a set E

are Input/Output **S**-equivalent (or simply **S**-equivalent) under a set of labellings $\mathfrak{L}' \subseteq \mathfrak{L}_E$ iff for any labelling $\mathbf{L}_E \in \mathfrak{L}'$ it holds that $\mathbf{S}\text{-}\mathtt{eff}_E(\mathscr{M}_1, \mathbf{L}_E) = \mathbf{S}\text{-}\mathtt{eff}_E(\mathscr{M}_2, \mathbf{L}_E)$. The multipoles $\mathscr{M}_1$ and $\mathscr{M}_2$ are **S**-equivalent iff they are **S**-equivalent under $\mathfrak{L}_E$.

It is easy to see that if two multipoles w.r.t. E are **S**-equivalent then they are **S**-equivalent under any set $\mathfrak{L}' \subseteq \mathfrak{L}_E$.

In Example 10, Example 11 and Example 14 $\mathscr{M}_1$ and $\mathscr{M}_2$ are **GR**-equivalent and **PR**-equivalent, while in Example 12 and Example 13 $\mathscr{M}_1$ and $\mathscr{M}_2$ are **PR**-equivalent but not **GR**-equivalent.

Replacements and transparent argumentation semantics

As anticipated by previous examples, an argumentation multipole can be viewed as a component of an argumentation framework that can be *replaced* with another multipole giving rise to a (possibly) different argumentation framework. In particular, given an argumentation framework $AF = (Ar, \rightarrow)$, one may partition the set of arguments *Args* into two sets, i.e. a set E which is not involved in the replacement and the set $D_1 = Ar \setminus E$ which is replaced along with the relevant attacks: the set D_1 identifies the multipole $\mathscr{M}_1 = (AF{\downarrow_{D_1}}, \rightarrow \cap (E \times D_1), \rightarrow \cap (D_1 \times E))$ w.r.t. E, which can be replaced with another multipole $\mathscr{M}_2$ w.r.t. the same set E. For later use in the chapter, it is worth identifying those replacements such that a partition belonging to the set returned by a selector $\mathscr{F}$ is enforced both before and after the replacement.

Definition 33. Let $AF = (Ar, \rightarrow)$ be an argumentation framework, and $E \subseteq Ar$ be a subset of its arguments. Let $D_1 = Ar \setminus E$, $R^1_{INP} = \rightarrow \cap (E \times D_1)$ and $R^1_{OUTP} = \rightarrow \cap (D_1 \times E)$. A *replacement* $\mathscr{R}$ is a tuple $(AF, \mathscr{M}_1, \mathscr{M}_2)$ where $\mathscr{M}_1 = (AF{\downarrow_{D_1}}, R^1_{INP}, R^1_{OUTP})$ and $\mathscr{M}_2$ is an argumentation multipole w.r.t. E. The set E is called the *invariant set* of the replacement $\mathscr{R}$. Assuming $\mathscr{M}_2 = ((D_2, R_{D_2}), R^2_{INP}, R^2_{OUTP})$, the result of the replacement $\mathscr{R}$, denoted as $T(\mathscr{R})$, is the argumentation framework $AF_2 = (E \cup D_2, (\rightarrow \cap E \times E) \cup R^2_{INP} \cup R_{D_2} \cup R^2_{OUTP})$. Given a partition selector $\mathscr{F}$, a replacement $(AF, \mathscr{M}_1, \mathscr{M}_2)$ is $\mathscr{F}$-preserving if both $(\{E, D_1\} \setminus \emptyset) \in \mathscr{F}(AF)$ and $(\{E, D_2\} \setminus \emptyset) \in \mathscr{F}(T(AF, \mathscr{M}_1, \mathscr{M}_2))$.

It is easy to see that $T(AF, \mathscr{M}_1, \mathscr{M}_1) = AF$. Moreover, letting $AF_2 = T(AF, \mathscr{M}_1, \mathscr{M}_2)$ it holds that $T(AF_2, \mathscr{M}_2, \mathscr{M}_1) = AF$. Note that, in the definition of $\mathscr{F}$-preserving replacement, the empty set is excluded from the requirement of belonging to $\mathscr{F}(AF)$. The reason is that by definition the empty set does not belong to any partition, however in case one of the sets in $\{E, D_1\}$ or $\{E, D_2\}$ is empty then it is sensible to require only the nonempty set to belong to $\mathscr{F}(AF)$.

[13] The reader may wonder why this condition has never been considered in the context of decomposability properties. The reason is that decomposability refers to partitions of the argumentation framework, which by definition include nonempty sets only.

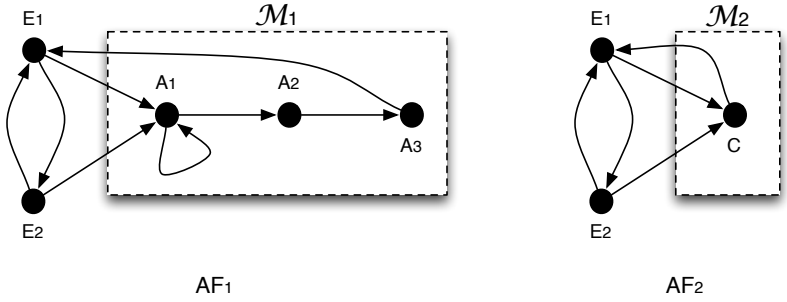


Figure 2.13: A contextually **PR**-legitimate replacement (Examples 15 and 17).

In Examples 10–14, the result of the replacement $(AF_1, \mathscr{M}_1, \mathscr{M}_2)$ is the argumentation framework AF_2.

While Definition 33 leaves room for any possible replacement, not all of them can be considered legitimate. In particular, we seek for replacements involving multipoles having the same Input/Output behavior, otherwise in most cases the labellings of the resulting frameworks would be different in the invariant set E, leading to changes in the status assignment of the relevant arguments. For instance, in Example 10 replacing $\mathscr{M}_1$ in AF_1 with a multipole including a single argument (or an odd-length chain of arguments) would change the label assigned to E_2 from `out` to `in`. In order to explore the notion of legitimate replacements, let us consider an issue arising e.g. in the following example.

Example 15. Consider the application of preferred semantics on the argumentation frameworks AF_1 and AF_2 shown in Figure 2.13, where $\mathscr{M}_1 = (AF_1{\downarrow}_{\{A_1,A_2,A_3\}}, \{(E_1,A_1),(E_2,A_1)\}, \{(A_3,E_1)\})$ and $\mathscr{M}_2 = (AF_2{\downarrow}_{\{C\}}, \{(E_1,C),(E_2,C)\}, \{(C,E_1)\})$ are two argumentation multipoles w.r.t. $\{E_1,E_2\}$, $AF_2 = T(\mathscr{R})$ with $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$, and the invariant set of the replacement $\mathscr{R}$ is $E = \{E_1,E_2\}$. The multipole $\mathscr{M}_1$ is *not* **PR**-equivalent to $\mathscr{M}_2$: considering the labelling $\{(E_1,\texttt{out}),(E_2,\texttt{out})\}$ $F_{\mathbf{PR}}$ prescribes for $\mathscr{M}_1$ the unique labelling $\{(A_1,\texttt{undec}),(A_2,\texttt{undec}),(A_3,\texttt{undec})\}$, whose effect on $\{E_1,E_2\}$ is $\{(E_1,\texttt{undec}),(E_2,\texttt{in})\}$, while $F_{\mathbf{PR}}$ prescribes for $\mathscr{M}_2$ the unique labelling $\{(C,\texttt{in})\}$, whose effect on $\{E_1,E_2\}$ is $\{(E_1,\texttt{out}),(E_2,\texttt{in})\}$. However, taking into account the possible labellings of $\mathscr{M}_1$ and $\mathscr{M}_2$, it can be noted that the labelling $\{(E_1,\texttt{out}),(E_2,\texttt{out})\}$ is impossible both in AF_1 and in AF_2. As to AF_1, if A_3 is `in` then $F_{\mathbf{PR}}$ prescribes for $\{E_1,E_2\}$ the labelling $\{(E_1,\texttt{out}),(E_2,\texttt{in})\}$, if A_3 is `out` then it prescribes the labellings $\{(E_1,\texttt{out}),(E_2,\texttt{in})\}$ and $\{(E_1,\texttt{in}),(E_2,\texttt{out})\}$, if A_3 is `undec` then it prescribes the labelling $\{(E_1,\texttt{out}),(E_2,\texttt{in})\}$. As to AF_2, the situation is the same. Summing up, the set of labellings that can be "seen" by $\mathscr{M}_1$ and $\mathscr{M}_2$ is $\mathfrak{L}^{\mathbf{PR}}_{\mathscr{R}} = \{\{(E_1,\texttt{out}),(E_2,\texttt{in})\},\{(E_1,\texttt{in}),(E_2,\texttt{out})\}\}$, under which $\mathscr{M}_1$ and $\mathscr{M}_2$ turn out to be **PR**-equivalent. In fact, for each of the labellings in $\mathfrak{L}^{\mathbf{PR}}_{\mathscr{R}}$, $F_{\mathbf{PR}}$ prescribes for $\mathscr{M}_1$ the unique labelling $\{(A_1,\texttt{out}),(A_2,\texttt{in}),(A_3,\texttt{out})\}$, whose effect on $\{E_1,E_2\}$ is $\{(E_1,\texttt{in}),(E_2,\texttt{in})\}$, and $F_{\mathbf{PR}}$ prescribes for $\mathscr{M}_2$ the unique labelling $\{(C,\texttt{out})\}$, whose effect on $\{E_1,E_2\}$ is again $\{(E_1,\texttt{in}),(E_2,\texttt{in})\}$.

Thus, a replacement may be considered as legitimate even if the involved multipoles are not equivalent under all labellings, provided that they are equivalent under the *possible* ones (in a sense, input labellings that never occur are neglected as the "don't care terms" in digital logic). Of course, one may accept to replace a multipole only with an equivalent one, since in this case equivalence holds independently of the context (in particular, the multipoles would remain equivalent even modifying the attack relations between arguments of the invariant set E). In order to distinguish between the two cases, a replacement is called

contextually legitimate in the first case, and simply *legitimate* in the latter. Independently of its legitimacy properties, we call *safe* a replacement that does not yield modifications of the labellings in E.

Definition 34. Let $\mathbf{S}$ be an argumentation semantics and $AF = (Ar, \rightarrow)$ be an argumentation framework. A replacement $\mathscr{R} = (AF, \mathscr{M}_1, \mathscr{M}_2)$ with invariant set E is $\mathbf{S}$-*legitimate* if $\mathscr{M}_1$ and $\mathscr{M}_2$ are $\mathbf{S}$-equivalent, it is *contextually* $\mathbf{S}$-*legitimate* if $\mathscr{M}_1$ and $\mathscr{M}_2$ are $\mathbf{S}$-equivalent under $\mathfrak{L}^{\mathbf{S}}_{\mathscr{R}}$, where $\mathfrak{L}^{\mathbf{S}}_{\mathscr{R}} \triangleq \{F_{\mathbf{S}}(AF{\downarrow_E}, {\mathscr{M}_1}^{\text{outp}}, \mathbf{L}_1, R^1_{OUTP}) \mid \mathbf{L}_1 \in \mathfrak{L}_{\mathscr{M}_1\text{outp}}\} \cup \{F_{\mathbf{S}}(AF{\downarrow_E}, {\mathscr{M}_2}^{\text{outp}}, \mathbf{L}_2, R^2_{OUTP}) \mid \mathbf{L}_2 \in \mathfrak{L}_{\mathscr{M}_2\text{outp}}\}$. Moreover, $\mathscr{R}$ is $\mathbf{S}$-*safe* if $\{\mathbf{L}{\downarrow_E} \mid \mathbf{L} \in \mathbf{L_S}(AF)\} = \{\mathbf{L}{\downarrow_E} \mid \mathbf{L} \in \mathbf{L_S}(T(AF, \mathscr{M}_1, \mathscr{M}_2))\}$.

It is easy to see that every legitimate replacement is also contextually legitimate. For instance, in Example 12 the replacement $(AF_1, \mathscr{M}_1, \mathscr{M}_2)$ is **PR**-legitimate and **PR**-safe, it is not contextually **GR**-legitimate nor **GR**-safe. In Example 15 the replacement $(AF_1, \mathscr{M}_1, \mathscr{M}_2)$ is contextually **PR**-legitimate (but not **PR**-legitimate) and **PR**-safe, and the same holds according to grounded semantics.

The examples presented so far may give the impression that for any semantics $\mathbf{S}$ a (possibly contextually) $\mathbf{S}$-legitimate replacement is always $\mathbf{S}$-safe, i.e. replacing a multipole with an equivalent multipole preserves the labellings in the invariant set of the replacement. This property may seem natural and easy to prove, however it is shown in Section 2.9 that it does not hold for all semantics: we denote as *transparent* the semantics such that legitimate replacements are always safe, *strongly transparent* the semantics such that contextually legitimate replacements are always safe. Similarly to decomposability, also transparency may hold under a restriction on the partition identified by the multipoles that are replaced: accordingly, we introduce the concept of transparency w.r.t. a partition selector $\mathscr{F}$.

Definition 35. A semantics $\mathbf{S}$ is *transparent* if any $\mathbf{S}$-legitimate replacement is $\mathbf{S}$-safe, it is *strongly transparent* if any contextually $\mathbf{S}$-legitimate replacement is $\mathbf{S}$-safe. Given a partition selector $\mathscr{F}$, a semantics $\mathbf{S}$ is transparent w.r.t. $\mathscr{F}$ if any $\mathscr{F}$-preserving and $\mathbf{S}$-legitimate replacement is $\mathbf{S}$-safe, it is strongly transparent w.r.t. $\mathscr{F}$ if any $\mathscr{F}$-preserving and contextually $\mathbf{S}$-legitimate replacement is $\mathbf{S}$-safe.

Since any ($\mathscr{F}$-preserving) legitimate replacement is also contextually legitimate, any strongly transparent semantics (w.r.t. $\mathscr{F}$) is also transparent (w.r.t. $\mathscr{F}$).

A limit case which is theoretically interesting to consider is a replacement $(AF, \mathscr{M}_1, \mathscr{M}_2)$ with the invariant set E equal to the empty set, i.e. when an entire argumentation framework is replaced by another one.

Proposition 12. Consider a semantics $\mathbf{S}$ such that $\mathbf{L_S}(AF_\emptyset) = \{\emptyset\}$ and a replacement $\mathscr{R} = (AF, \mathscr{M}_1, \mathscr{M}_2)$ with invariant set $E = \emptyset$. Letting $AF_2 = T(\mathscr{R})$, the following conditions are equivalent:

- $\mathscr{R}$ is $\mathbf{S}$-legitimate
- $\mathscr{R}$ is contextually $\mathbf{S}$-legitimate
- $|\mathbf{L_S}(AF)| > 0 \wedge |\mathbf{L_S}(AF_2)| > 0$, or $\mathbf{L_S}(AF) = \mathbf{L_S}(AF_2) = \emptyset$
- $\mathscr{R}$ is $\mathbf{S}$-safe.

Intuitively, there are no preserved arguments, thus the effect of any labelling of AF on the outside empty set is the same as the effect of any labelling of AF_2. The only difference arises in the case that AF "crashes" (i.e. admits no labellings) while AF_2 does not exhibit such pathological behavior, or vice versa.

Note that the notions of replacement and transparent semantics refer to partitions of argumentation frameworks into just two subframeworks, i.e. one corresponding to the replaced multipole $\mathscr{M}_1$ (or the replacing one $\mathscr{M}_2$) and the other identified by the invariant set E. This is not restrictive, since one can treat a multiple replacement of several multipoles as a sequence of replacements each involving just one multipole. The following proposition shows that safeness is preserved by a sequence of safe replacements, and the same holds for skeptical and credulous justification of those arguments that are not replaced.

Proposition 13. Let $AF = (Ar, \rightarrow)$ be an argumentation framework. Consider a sequence of replacements $(R_1, R_2, \ldots, R_n)$ where $R_i = (AF_i, \mathscr{M}_{i,1}, \mathscr{M}_{i,2})$, E_i is the invariant set of R_i, $AF_1 = AF$ and, for any $1 < i \leq n$, $AF_i = T(AF_{i-1}, \mathscr{M}_{i-1,1}, \mathscr{M}_{i-1,2})$. Let AF_* be the result of the sequence of replacements, i.e. $AF_* = T(AF_n, \mathscr{M}_{n,1}, \mathscr{M}_{n,2})$. If all replacements R_i are **S**-safe, then letting $E = E_1 \cap \cdots \cap E_n$ it holds that $\{\mathbf{L}{\downarrow_E} \mid \mathbf{L} \in \mathbf{L_S}(AF)\} = \{\mathbf{L}{\downarrow_E} \mid \mathbf{L} \in \mathbf{L_S}(AF_*)\}$. Moreover, any argument $A \in E$ is skeptically/credulously justified according to **S** in AF if and only if it is skeptically/credulously justified according to **S** in AF_*.

2.8 The relationship between decomposability and transparency

Intuitively, there is a close relationship between decomposability and transparency: if a semantics is decomposable, i.e. the labellings prescribed for an argumentation framework are completely determined by applying the canonical local function to the elements of a partition, then one may expect that replacing a multipole with another one having the same Input/Output behavior has no impact on the invariant set of the replacement. This intuition is confirmed by Theorem 8, showing that decomposability of a semantics **S** is a sufficient condition for strong transparency.

Theorem 8. *Consider an effect-dictated semantics* **S** *such that* $\mathbf{L_S}(AF_\emptyset) = \{\emptyset\}$. *If* **S** *is decomposable w.r.t. a partition selector* $\mathscr{F}$ *then* **S** *is strongly transparent w.r.t.* $\mathscr{F}$.

While full decomposability is a sufficient condition for strong transparency, it is not necessary. In particular, for a single-status semantics which is top-down decomposable a relaxed form of bottom-up decomposability is sufficient to ensure strong transparency. More specifically, in this case bottom-up decomposability requires the union of local labellings to coincide with the (unique) global labelling. However, just requiring the union of local labellings to be more or equally committed than the global labelling is enough to achieve strong transparency, as shown by Theorem 9.

Theorem 9. *Let* **S** *be an effect-dictated single-status semantics such that* $\mathbf{L_S}(AF_\emptyset) = \{\emptyset\}$. *Suppose that* **S** *is top-down decomposable w.r.t. a partition selector* $\mathscr{F}$ *and satisfies the following property: for any argumentation framework AF and any partition* $\{E, D\} \in \mathscr{F}(AF)$, *letting* **L** *be the labelling prescribed by* **S** *for AF, if* $\mathbf{L}^E \in \mathfrak{L}_E$ *and* $\mathbf{L}^D \in \mathfrak{L}_D$ *are two labellings such that* $\mathbf{L}^E \in F_\mathbf{S}(AF{\downarrow_E}, E^{inp}, \mathbf{L}^D{\downarrow_{E^{inp}}}, E^R)$ *and* $\mathbf{L}^D \in F_\mathbf{S}(AF{\downarrow_D}, D^{inp}, \mathbf{L}^E{\downarrow_{D^{inp}}}, D^R)$, *then* $\mathbf{L} \sqsubseteq \mathbf{L}^E \cup \mathbf{L}^D$. *Then* **S** *is strongly transparent w.r.t.* $\mathscr{F}$.

2.9 Analyzing transparency of argumentation semantics

In this section we discuss the transparency properties of the semantics reviewed in Section 2.3. A synthetic view of the results is given in Table 2.2 (for all semantics strong transparency turns out to be equivalent to transparency, and any transparency property w.r.t. $\mathscr{F}_{\cup\mathsf{SCC}}$ holds if and only if the same property holds w.r.t. $\mathscr{F}_{\mathsf{SCC}}$).

	AD	CO	ST	GR	PR	ID	SST
(Strong) transparency	Yes	Yes	Yes	Yes	No*	No	No
(Strong) transparency w.r.t. $\mathscr{F}_{\cup \mathsf{SCC}}$ and $\mathscr{F}_{\mathsf{SCC}}$	Yes	Yes	Yes	Yes	Yes	No	No
(Strong) transparency in case of acyclic multipoles	Yes	Yes	Yes	Yes	Yes	Yes	No

* holds under additional conditions (see Definitions 36 and 37).

Table 2.2: Transparency properties of argumentation semantics.

Admissible, complete and stable semantics

As shown in Section 2.5, admissible, complete and stable semantics satisfy full decomposability: this easily yields strong transparency for such semantics.

Theorem 10. *Admissible semantics* **AD***, complete semantics* **CO** *and stable semantics* **ST** *are strongly transparent.*

For instance, in Examples 10 and 11 the replacement $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$ is **S**-legitimate, where $\mathbf{S} \in \{\mathbf{AD}, \mathbf{CO}, \mathbf{ST}\}$, therefore it is also **S**-safe, i.e. $\{\mathbf{L}{\downarrow_E} \mid \mathbf{L} \in \mathbf{L_S}(AF_1)\} = \{\mathbf{L}{\downarrow_E} \mid \mathbf{L} \in \mathbf{L_S}(AF_2)\}$. In Examples 12, 13 and 14 $\mathscr{R}$ is **ST**-legitimate, in Example 15 it is contextually **ST**-legitimate, therefore in all cases $\mathscr{R}$ is **ST**-safe. In particular, in Example 15 $\mathbf{L_{ST}}(AF_1) = \{\{(E_1, \mathtt{in}), (E_2, \mathtt{out}), (A_1, \mathtt{out}), (A_2, \mathtt{in}), (A_3, \mathtt{out})\}, \{(E_1, \mathtt{out}), (E_2, \mathtt{in}), (A_1, \mathtt{out}), (A_2, \mathtt{in}), (A_3, \mathtt{out})\}\}$ and $\mathbf{L_{ST}}(AF_2) = \{\{(E_1, \mathtt{in}), (E_2, \mathtt{out}), (C, \mathtt{out})\}, \{(E_1, \mathtt{out}), (E_2, \mathtt{in}), (C, \mathtt{out})\}\}$, thus the stable labellings restricted to $\{E_1, E_2\}$ are $\{(E_1, \mathtt{out}), (E_2, \mathtt{in})\}$ and $\{(E_1, \mathtt{in}), (E_2, \mathtt{out})\}$ both in AF_1 and in AF_2.

Grounded semantics

As shown in Section 2.5, grounded semantics is not fully decomposable but only top-down decomposable.

Theorem 11 shows however that grounded semantics is strongly transparent, building on the result proved in Theorem 9.

Theorem 11. *Grounded semantics* **GR** *is strongly transparent.*

For instance, in Examples 10, 11 and 14 the replacement $(AF_1, \mathscr{M}_1, \mathscr{M}_2)$ is **GR**-legitimate, therefore it is also **GR**-safe. In Example 15 the replacement $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$ is contextually **GR**-legitimate, since $\mathfrak{L}^{\mathbf{GR}}_{\mathscr{R}} = \{\{(E_1, \mathtt{undec}), (E_2, \mathtt{undec})\}, \{(E_1, \mathtt{out}), (E_2, \mathtt{in})\}\}$ and $\mathscr{M}_1$ and $\mathscr{M}_2$ are **GR**-equivalent under $\mathfrak{L}^{\mathbf{GR}}_{\mathscr{R}}$: as a consequence, $\mathscr{R}$ is **GR**-safe, as it can be seen by considering that both the grounded labelling of AF_1 and the grounded labelling of AF_2 assign to all arguments the label `undec`.

Preferred semantics

Like grounded semantics, preferred semantics is top-down decomposable but not fully decomposable. However, differently from grounded semantics, preferred semantics is not transparent, as shown by the following counterexample.

Example 16. Consider the argumentation frameworks AF_1 and AF_2 shown in Figure 2.14, where $AF_2 = T(\mathscr{R})$ with $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$, and the invariant set of the replacement $\mathscr{R}$ is $E = \{E_1, E_2\}$. It turns out

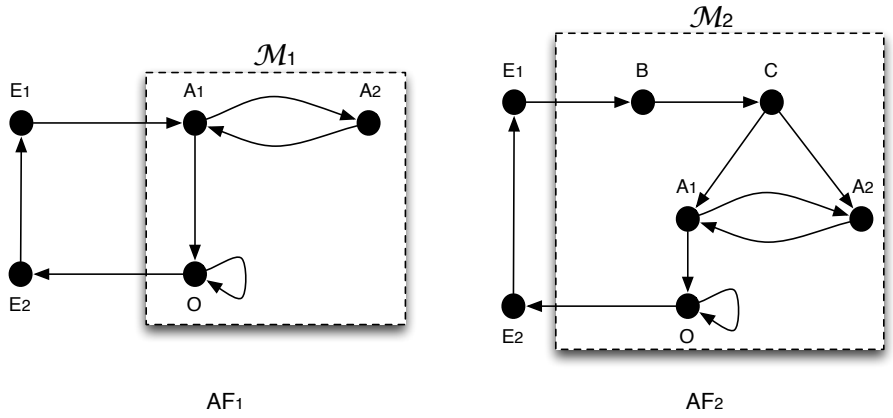


Figure 2.14: Preferred semantics is not transparent (Examples 16 and 18).

that $\mathscr{M}_1$ and $\mathscr{M}_2$ are **PR**-equivalent, thus $\mathscr{R}$ is **PR**-legitimate. In fact, for any label $\mathbf{L}^{in} \in \mathfrak{L}_E$ such that E_1 is labelled `in` the local function $F_{\mathbf{PR}}$ prescribes for $\mathscr{M}_1$ the unique labelling $\{(A_1,\texttt{out}),(A_2,\texttt{in}),(O,\texttt{undec})\}$, therefore $\mathbf{PR}\text{-}\texttt{eff}_E(\mathscr{M}_1,\mathbf{L}^{in}) = \{\{(E_2,\texttt{undec}),(E_1,\texttt{in})\}\}$, and it prescribes for $\mathscr{M}_2$ the unique labelling $\{(B,\texttt{out}),(C,\texttt{in}),(A_1,\texttt{out}),(A_2,\texttt{out}),(O,\texttt{undec})\}$, therefore also $\mathbf{PR}\text{-}\texttt{eff}_E(\mathscr{M}_2,\mathbf{L}^{in}) = \{\{(E_2,\texttt{undec}),(E_1,\texttt{in})\}\}$. For any label $\mathbf{L}^{out} \in \mathfrak{L}_E$ such that E_1 is labelled `out` $F_{\mathbf{PR}}$ prescribes for $\mathscr{M}_1$ the labellings $\{(A_1,\texttt{in}),(A_2,\texttt{out}),(O,\texttt{out})\}$ and $\{(A_1,\texttt{out}),(A_2,\texttt{in}),(O,\texttt{undec})\}$, for $\mathscr{M}_2$ the labellings $\{(B,\texttt{in}),(C,\texttt{out}),(A_1,\texttt{in}),(A_2,\texttt{out}),(O,\texttt{out})\}$ and $\{(B,\texttt{in}),(C,\texttt{out}),(A_1,\texttt{out}),(A_2,\texttt{in}),(O,\texttt{undec})\}$, thus $\mathbf{PR}\text{-}\texttt{eff}_E(\mathscr{M}_1,\mathbf{L}^{out}) = \mathbf{PR}\text{-}\texttt{eff}_E(\mathscr{M}_2,\mathbf{L}^{out}) = \{\{(E_2,\texttt{in}),(E_1,\texttt{in})\},\{(E_2,\texttt{undec}),(E_1,\texttt{in})\}\}$. For any label $\mathbf{L}^{undec} \in \mathfrak{L}_E$ such that E_1 is labelled `undec`, $F_{\mathbf{PR}}$ prescribes for $\mathscr{M}_1$ the unique labelling $\{(A_1,\texttt{out}),(A_2,\texttt{in}),(O,\texttt{undec})\}$, and it prescribes for $\mathscr{M}_2$ the unique labelling $\{(B,\texttt{undec}),(C,\texttt{undec}),(A_1,\texttt{undec}),(A_2,\texttt{undec}),(O,\texttt{undec})\}$, therefore $\mathbf{PR}\text{-}\texttt{eff}_E(\mathscr{M}_1,\mathbf{L}^{undec}) = \mathbf{PR}\text{-}\texttt{eff}_E(\mathscr{M}_2,\mathbf{L}^{undec}) = \{\{(E_2,\texttt{undec}),(E_1,\texttt{in})\}\}$. However, the replacement $(AF_1,\mathscr{M}_1,\mathscr{M}_2)$ is not **PR**-safe. In fact, the preferred labellings of AF_1 are $\{(A_1,\texttt{in}),(A_2,\texttt{out}),(O,\texttt{out}),(E_2,\texttt{in}),(E_1,\texttt{out})\}$ and $\{(A_1,\texttt{out}),(A_2,\texttt{in}),(O,\texttt{undec}),(E_2,\texttt{undec}),(E_1,\texttt{undec})\}$, while $\{(B,\texttt{in}),(C,\texttt{out}),(A_1,\texttt{in}),(A_2,\texttt{out}),(O,\texttt{out}),(E_2,\texttt{in}),(E_1,\texttt{out})\}$ is the only preferred labelling of AF_2. Note in particular that E_2 is skeptically justified in AF_2 but not in AF_1.
Interestingly enough, considering the application of stable semantics it can be checked that the replacement $(AF_1,\mathscr{M}_1,\mathscr{M}_2)$ is **ST**-legitimate, therefore according to Theorem 10 it is also **ST**-safe. In fact, $\mathbf{L}_{\mathbf{ST}}(AF_1) = \{\{(A_1,\texttt{in}),(A_2,\texttt{out}),(O,\texttt{out}),(E_2,\texttt{in}),(E_1,\texttt{out})\}\}$ and $\mathbf{L}_{\mathbf{ST}}(AF_2) = \{\{(B,\texttt{in}),(C,\texttt{out}),(A_1,\texttt{in}),(A_2,\texttt{out}),(O,\texttt{out}),(E_2,\texttt{in}),(E_1,\texttt{out})\}\}$, therefore both in AF_1 and in AF_2 the argument E_1 is labelled `out` and E_2 is labelled `in` by all stable labellings.

In the previous example a **PR**-legitimate replacement yields a change in the status assignment of arguments belonging to the invariant set E, however it can be noted that their credulous justification is preserved, i.e. E_2 is credulously justified both in AF_1 and AF_2, E_1 is not credulously justified either in AF_1 or in AF_2. Theorem 12 proves that this result holds in general.

Theorem 12. *For any contextually* **PR***-legitimate replacement $\mathscr{R} = (AF,\mathscr{M}_1,\mathscr{M}_2)$ with invariant set E, any argument $A \in E$ is credulously justified according to* **PR** *in AF if and only if it is credulously justified according to* **PR** *in $T(AF,\mathscr{M}_1,\mathscr{M}_2)$.*

While the obtained result is somewhat weak, as it concerns credulous justification only, it has to be acknowledged that the counterexample against transparency of **PR** (Example 16) is rather tricky. In particular, $\mathscr{M}_1$ and $\mathscr{M}_2$ are **PR**-equivalent, but they differ in the following aspect. On the one hand, in $\mathscr{M}_1$, the local function $F_{\mathbf{PR}}$ prescribes for any input labelling $\mathbf{L}^{undec}$ the unique labelling $\{(A_1,\mathtt{out}),(A_2,\mathtt{in}),(O,\mathtt{undec})\}$, and with the "more committed" input labelling $\mathbf{L}^{out} \in \mathfrak{L}_E$ it returns (among others) the labelling $\{(A_1,\mathtt{in}),(A_2,\mathtt{out}),(O,\mathtt{out})\}$ which is not "more committed" than $\{(A_1,\mathtt{out}),(A_2,\mathtt{in}),(O,\mathtt{undec})\}$, i.e. it is not the case that $\{(A_1,\mathtt{out}),(A_2,\mathtt{in}),(O,\mathtt{undec})\} \sqsubseteq \{(A_1,\mathtt{in}),(A_2,\mathtt{out}),(O,\mathtt{out})\}$. On the other hand, in $\mathscr{M}_2$ both the labellings returned by the local function $F_{\mathbf{PR}}$ with the input labelling $\mathbf{L}^{out}$ are "more committed" than the labelling returned by $F_{\mathbf{PR}}$ with the input labelling $\mathbf{L}^{undec}$, i.e. it holds that $\{(B,\mathtt{undec}),(C,\mathtt{undec}),(A_1,\mathtt{undec}),(A_2,\mathtt{undec}),(O,\mathtt{undec})\} \sqsubseteq \{(B,\mathtt{in}),(C,\mathtt{out}),(A_1,\mathtt{in}),(A_2,\mathtt{out}),(O,\mathtt{out})\}$, $\{(B,\mathtt{undec}),(C,\mathtt{undec}),(A_1,\mathtt{undec}),(A_2,\mathtt{undec}),(O,\mathtt{undec})\} \sqsubseteq \{(B,\mathtt{in}),(C,\mathtt{out}),(A_1,\mathtt{out}),(A_2,\mathtt{in}),(O,\mathtt{undec})\}$. More generally, we define the notion of *homogeneously equivalent* argumentation multipoles, corresponding to equivalent multipoles that exhibit a sort of mutually regular behavior.

Definition 36. Two multipoles $\mathscr{M}_1 = (AF_1, R^1_{INP}, R^1_{OUTP})$ and $\mathscr{M}_2 = (AF_2, R^2_{INP}, R^2_{OUTP})$ w.r.t. a set E are homogeneously **S**-equivalent under a set of labellings $\mathfrak{L}' \subseteq \mathfrak{L}_E$ iff they are **S**-equivalent under $\mathfrak{L}'$ and the following two symmetric conditions hold:

1. Given $\mathbf{L}^1_E, \mathbf{L}^2_E \in \mathfrak{L}'$ such that $\mathbf{L}^1_E \sqsubseteq \mathbf{L}^2_E$,
   if there are two labellings $\mathbf{L}_1^{D_1} \in F_{\mathbf{S}}(AF_1, {\mathscr{M}_1}^{\mathrm{inp}}, \mathbf{L}^1_E{\downarrow}_{{\mathscr{M}_1}^{\mathrm{inp}}}, R^1_{INP})$ and $\mathbf{L}_2^{D_1} \in F_{\mathbf{S}}(AF_1, {\mathscr{M}_1}^{\mathrm{inp}}, \mathbf{L}^2_E{\downarrow}_{{\mathscr{M}_1}^{\mathrm{inp}}}, R^1_{INP})$ such that $\mathbf{L}_1^{D_1} \sqsubseteq \mathbf{L}_2^{D_1}$, then $\forall \mathbf{L}_1^{D_2} \in F_{\mathbf{S}}(AF_2, {\mathscr{M}_2}^{\mathrm{inp}}, \mathbf{L}^1_E{\downarrow}_{{\mathscr{M}_2}^{\mathrm{inp}}}, R^2_{INP})$ such that $\mathtt{eff}_E({\mathscr{M}_2}^{\mathrm{outp}}, \mathbf{L}_1^{D_2}{\downarrow}_{{\mathscr{M}_2}^{\mathrm{outp}}}, R^2_{OUTP}) = \mathtt{eff}_E({\mathscr{M}_1}^{\mathrm{outp}}, \mathbf{L}_1^{D_1}{\downarrow}_{{\mathscr{M}_1}^{\mathrm{outp}}}, R^1_{OUTP})$, there is a labelling $\mathbf{L}_2^{D_2} \in F_{\mathbf{S}}(AF_2, {\mathscr{M}_2}^{\mathrm{inp}}, \mathbf{L}^2_E{\downarrow}_{{\mathscr{M}_2}^{\mathrm{inp}}}, R^2_{INP})$ such that $\mathtt{eff}_E({\mathscr{M}_2}^{\mathrm{outp}}, \mathbf{L}_2^{D_2}{\downarrow}_{{\mathscr{M}_2}^{\mathrm{outp}}}, R^2_{OUTP}) = \mathtt{eff}_E({\mathscr{M}_1}^{\mathrm{outp}}, \mathbf{L}_2^{D_1}{\downarrow}_{{\mathscr{M}_1}^{\mathrm{outp}}}, R^1_{OUTP})$ and $\mathbf{L}_1^{D_2} \sqsubseteq \mathbf{L}_2^{D_2}$.

2. Given $\mathbf{L}^1_E, \mathbf{L}^2_E \in \mathfrak{L}'$ such that $\mathbf{L}^1_E \sqsubseteq \mathbf{L}^2_E$, if there are two labellings $\mathbf{L}_1^{D_2} \in F_{\mathbf{S}}(AF_2, {\mathscr{M}_2}^{\mathrm{inp}}, \mathbf{L}^1_E{\downarrow}_{{\mathscr{M}_2}^{\mathrm{inp}}}, R^2_{INP})$ and $\mathbf{L}_2^{D_2} \in F_{\mathbf{S}}(AF_2, {\mathscr{M}_2}^{\mathrm{inp}}, \mathbf{L}^2_E{\downarrow}_{{\mathscr{M}_2}^{\mathrm{inp}}}, R^2_{INP})$ such that $\mathbf{L}_1^{D_2} \sqsubseteq \mathbf{L}_2^{D_2}$, then $\forall \mathbf{L}_1^{D_1} \in F_{\mathbf{S}}(AF_1, {\mathscr{M}_1}^{\mathrm{inp}}, \mathbf{L}^1_E{\downarrow}_{{\mathscr{M}_1}^{\mathrm{inp}}}, R^1_{INP})$ such that $\mathtt{eff}_E({\mathscr{M}_1}^{\mathrm{outp}}, \mathbf{L}_1^{D_1}{\downarrow}_{{\mathscr{M}_1}^{\mathrm{outp}}}, R^1_{OUTP}) = \mathtt{eff}_E({\mathscr{M}_2}^{\mathrm{outp}}, \mathbf{L}_1^{D_2}{\downarrow}_{{\mathscr{M}_2}^{\mathrm{outp}}}, R^2_{OUTP})$, there is a labelling $\mathbf{L}_2^{D_1} \in F_{\mathbf{S}}(AF_1, {\mathscr{M}_1}^{\mathrm{inp}}, \mathbf{L}^2_E{\downarrow}_{{\mathscr{M}_1}^{\mathrm{inp}}}, R^1_{INP})$ such that $\mathtt{eff}_E({\mathscr{M}_1}^{\mathrm{outp}}, \mathbf{L}_2^{D_1}{\downarrow}_{{\mathscr{M}_1}^{\mathrm{outp}}}, R^1_{OUTP}) = \mathtt{eff}_E({\mathscr{M}_2}^{\mathrm{outp}}, \mathbf{L}_2^{D_2}{\downarrow}_{{\mathscr{M}_2}^{\mathrm{outp}}}, R^2_{OUTP})$ and $\mathbf{L}_1^{D_1} \sqsubseteq \mathbf{L}_2^{D_1}$.

In Example 16, it can be seen that the argumentation multipoles $\mathscr{M}_1$ and $\mathscr{M}_2$, while being **PR**-equivalent, are not homogeneously **PR**-equivalent.

It turns out that strong transparency of preferred semantics is recovered in case of replacements involving homogeneously **PR**-equivalent multipoles.

Theorem 13. *Any replacement $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$ with invariant set E, such that $\mathscr{M}_1$ and $\mathscr{M}_2$ are homogeneously* **PR***-equivalent under $\mathfrak{L}^{\mathbf{PR}}_{\mathscr{R}}$, is* **PR***-safe.*

Given two equivalent multipoles, a sufficient condition for their homogeneous equivalence is that each multipole is "internally homogeneous", i.e. the labellings prescribed by the local function are related by set-inclusion in a regular way w.r.t. the commitment relation between the input labellings. Definition 37 formalizes this intuition, while the sufficiency result is proved by Lemma 5 and Corollary 1.

Definition 37. Consider an argumentation semantics **S**. An argumentation multipole $\mathscr{M} = (AF, R_{INP}, R_{OUTP})$ w.r.t. a set E is *internally* **S**-*homogeneous* under a set of labellings $\mathfrak{L}' \subseteq \mathfrak{L}_E$ iff for all labellings $\mathbf{L}_E^1, \mathbf{L}_E^2 \in \mathfrak{L}'$ such that $\mathbf{L}_E^1 \sqsubseteq \mathbf{L}_E^2$, it holds that $\forall \mathbf{L}_1 \in F_{\mathbf{S}}(AF, \mathscr{M}^{\text{inp}}, \mathbf{L}_E^1{\downarrow}_{\mathscr{M}^{\text{inp}}}, R_{INP})$, $\forall \mathbf{L}_2 \in F_{\mathbf{S}}(AF, \mathscr{M}^{\text{inp}}, \mathbf{L}_E^2{\downarrow}_{\mathscr{M}^{\text{inp}}}, R_{INP})$ such that $\mathtt{eff}_E(\mathscr{M}^{\text{outp}}, \mathbf{L}_1{\downarrow}_{\mathscr{M}^{\text{outp}}}, R_{OUTP}) \sqsubseteq \mathtt{eff}_E(\mathscr{M}^{\text{outp}}, \mathbf{L}_2{\downarrow}_{\mathscr{M}^{\text{outp}}}, R_{OUTP})$, there is a labelling $\mathbf{L}_2' \in F_{\mathbf{S}}(AF, \mathscr{M}^{\text{inp}}, \mathbf{L}_E^2{\downarrow}_{\mathscr{M}^{\text{inp}}}, R_{INP})$ such that $\mathtt{eff}_E(\mathscr{M}^{\text{outp}}, \mathbf{L}_2{\downarrow}_{\mathscr{M}^{\text{outp}}}, R_{OUTP}) = \mathtt{eff}_E(\mathscr{M}^{\text{outp}}, \mathbf{L}_2'{\downarrow}_{\mathscr{M}^{\text{outp}}}, R_{OUTP})$ and $\mathbf{L}_1 \sqsubseteq \mathbf{L}_2'$.

Lemma 5. Consider two multipoles $\mathscr{M}_1 = (AF_1, R_{INP}^1, R_{OUTP}^1)$ and $\mathscr{M}_2 = (AF_2, R_{INP}^2, R_{OUTP}^2)$ w.r.t. a set E which are internally **S**-homogeneous under a set of labellings $\mathfrak{L}' \subseteq \mathfrak{L}_E$. If $\mathscr{M}_1$ and $\mathscr{M}_2$ are **S**-equivalent under $\mathfrak{L}'$, then they are homogeneously **S**-equivalent under $\mathfrak{L}'$.

Corollary 1. Any replacement $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$ with invariant set E, such that $\mathscr{M}_1$ and $\mathscr{M}_2$ are **PR**-equivalent under $\mathfrak{L}_{\mathscr{R}}^{\mathbf{PR}}$ and both $\mathscr{M}_1$ and $\mathscr{M}_2$ are internally **PR**-homogeneous under $\mathfrak{L}_{\mathscr{R}}^{\mathbf{PR}}$, is **PR**-safe.

Example 17. Consider again the replacement $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$ depicted in Figure 2.13. As shown in Example 15, $\mathfrak{L}_{\mathscr{R}}^{\mathbf{PR}} = \{\{(E_1, \mathtt{out}), (E_2, \mathtt{in})\}, \{(E_1, \mathtt{in}), (E_2, \mathtt{out})\}\}$ and $\mathscr{M}_1$, $\mathscr{M}_2$ are **PR**-equivalent under $\mathfrak{L}_{\mathscr{R}}^{\mathbf{PR}}$. Since there are no distinct labellings $\mathbf{L}_E^1, \mathbf{L}_E^2 \in \mathfrak{L}_{\mathscr{R}}^{\mathbf{PR}}$ such that $\mathbf{L}_E^1 \sqsubseteq \mathbf{L}_E^2$, $\mathscr{M}_1$ and $\mathscr{M}_2$ are trivially internally **PR**-homogeneous under $\mathfrak{L}_{\mathscr{R}}^{\mathbf{PR}}$. As a consequence, by Corollary 1 the replacement $\mathscr{R}$ is **PR**-safe. In fact, there are two preferred labellings in AF_1, namely $\{(E_1, \mathtt{in}), (E_2, \mathtt{out}), (A_1, \mathtt{out}), (A_2, \mathtt{in}), (A_3, \mathtt{out})\}$ and $\{(E_1, \mathtt{out}), (E_2, \mathtt{in}), (A_1, \mathtt{out}), (A_2, \mathtt{in}), (A_3, \mathtt{out})\}$, while in AF_2 the preferred labellings are $\{(E_1, \mathtt{in}), (E_2, \mathtt{out}), (C, \mathtt{out})\}$ and $\{(E_1, \mathtt{out}), (E_2, \mathtt{in}), (C, \mathtt{out})\}$. Thus, the restriction of the preferred labellings to $\{E_1, E_2\}$ are $\{(E_1, \mathtt{out}), (E_2, \mathtt{in})\}$ and $\{(E_1, \mathtt{in}), (E_2, \mathtt{out})\}$ both in AF_1 and in AF_2.

Turning to non arbitrary partitionings, strong transparency of preferred semantics is recovered without additional conditions for replacements involving the union of strongly connected components.

Theorem 14. *Preferred semantics* **PR** *is strongly transparent w.r.t.* $\mathscr{F}_{\cup\mathsf{SCC}}$.

Example 18. The multipoles $\mathscr{M}_1$ and $\mathscr{M}_2$ shown in Figure 2.14 can be safely interchanged if they correspond to the union of strongly connected components. For instance, removing the attack from E_2 to E_1 makes the replacement $(AF_1, \mathscr{M}_1, \mathscr{M}_2)$ $\mathscr{F}_{\cup\mathsf{SCC}}$-preserving, thus such replacement is safe. In fact, in this case there is a unique preferred labelling in AF_1 and a unique preferred labelling in AF_2, and in both cases E_1 is labelled `in` and E_2 is labelled `undec`.

It is easy to see that in Examples 10, 11, 12, 13 and 14 the replacement $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$ is $\mathscr{F}_{\cup\mathsf{SCC}}$-preserving and **PR**-legitimate. As a consequence, in all cases the replacement $\mathscr{R}$ is safe, i.e. $\{\mathbf{L}{\downarrow}_E \mid \mathbf{L} \in \mathbf{L}_{\mathbf{PR}}(AF_1)\} = \{\mathbf{L}{\downarrow}_E \mid \mathbf{L} \in \mathbf{L}_{\mathbf{PR}}(AF_2)\}$. Moreover, it can be seen that in all cases the multipoles are internally **PR**-homogeneous, therefore they could be safely interchanged also in the context of non $\mathscr{F}_{\cup\mathsf{SCC}}$-preserving replacements.

Ideal semantics

The transparency properties of ideal semantics mirror the discouraging decomposability properties analyzed in Section 2.5: the following example, inspired by Example 7, shows that ideal semantics is not transparent even w.r.t. $\mathscr{F}_{\mathsf{SCC}}$.

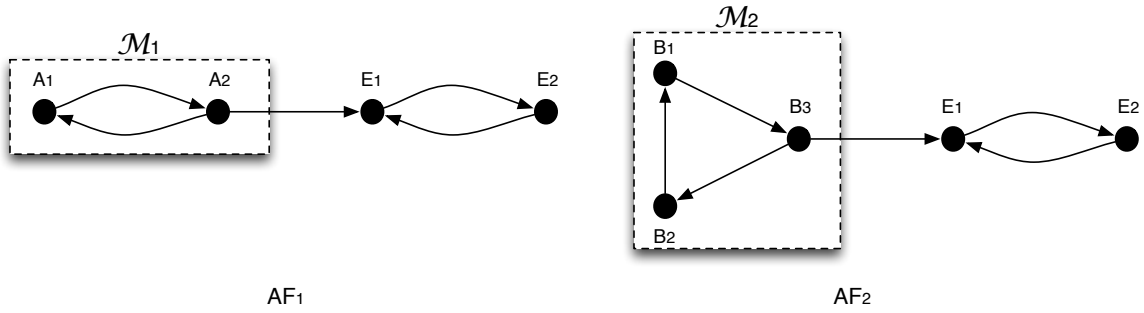


Figure 2.15: Ideal semantics is not transparent w.r.t. $\mathscr{F}_{\mathsf{SCC}}$ (Example 19).

Example 19. Consider the argumentation frameworks AF_1 and AF_2 shown in Figure 2.15, where $AF_2 = T(\mathscr{R})$ with $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$, and the invariant set of the replacement $\mathscr{R}$ is $E = \{E_1, E_2\}$. It is easy to see that $\mathscr{M}_1$ and $\mathscr{M}_2$ are **ID**-equivalent, since $F_{\mathbf{ID}}$ prescribes for $\mathscr{M}_1$ the labelling $\{(A_1, \mathtt{undec}), (A_2, \mathtt{undec})\}$ and for $\mathscr{M}_2$ the labelling $\{(B_1, \mathtt{undec}), (B_2, \mathtt{undec}), (B_3, \mathtt{undec})\}$. As a consequence, the replacement $\mathscr{R}$ is **ID**-legitimate, and it is also easy to see that it is $\mathscr{F}_{\mathsf{SCC}}$-preserving. However, $\mathscr{R}$ is not **ID**-safe, since the ideal labelling of AF_1 leaves all the arguments undec, while the ideal labelling of AF_2 is $\{(B_1, \mathtt{undec}), (B_2, \mathtt{undec}), (B_3, \mathtt{undec}), (E_1, \mathtt{out}), (E_2, \mathtt{in})\}$.

Transparency is recovered in the (somewhat specific) case of replacements involving multipoles for which $F_{\mathbf{CO}}$ always prescribes a unique labelling.

Definition 38. Consider an argumentation semantics **S**. An argumentation multipole $\mathscr{M} = (AF, R_{INP}, R_{OUTP})$ w.r.t. a set E is **S**-univocal under a set of labellings $\mathfrak{L}' \subseteq \mathfrak{L}_E$ iff $\forall \mathbf{L}_E \in \mathfrak{L}'$ $|F_{\mathbf{S}}(AF, \mathscr{M}^{\mathrm{inp}}, \mathbf{L}_E{\downarrow}_{\mathscr{M}^{\mathrm{inp}}}, R_{INP})| = 1$.

The following lemmas prove some specific results holding in the case of **CO**-univocal argumentation multipoles.

Lemma 6. Let $\mathscr{M}$ be an argumentation multipole (AF, R_{INP}, R_{OUTP}) w.r.t. a set E which is **CO**- univocal under a set of labellings $\mathfrak{L}' \subseteq \mathfrak{L}_E$. Then $\forall \mathbf{L}_E \in \mathfrak{L}'$, $F_{\mathbf{CO}}(AF, \mathscr{M}^{\mathrm{inp}}, \mathbf{L}_E{\downarrow}_{\mathscr{M}^{\mathrm{inp}}}, R_{INP}) = F_{\mathbf{S}}(AF, \mathscr{M}^{\mathrm{inp}}, \mathbf{L}_E{\downarrow}_{\mathscr{M}^{\mathrm{inp}}}, R_{INP})$ for any $\mathbf{S} \in \{\mathbf{GR}, \mathbf{PR}, \mathbf{ID}, \mathbf{SST}\}$.

Lemma 7. Let $\mathscr{M}$ be an argumentation multipole (AF, R_{INP}, R_{OUTP}) w.r.t. a set E which is **CO**- univocal under a set of labellings $\mathfrak{L}' \subseteq \mathfrak{L}_E$, and let $\mathbf{L}_E^1, \mathbf{L}_E^2$ be two labellings of $\mathfrak{L}'$ such that $\mathbf{L}_E^1 \sqsubseteq \mathbf{L}_E^2$. Then, for any two labellings $\mathbf{L}_1, \mathbf{L}_2$ such that $\mathbf{L}_1 \in F_{\mathbf{PR}}(AF, \mathscr{M}^{\mathrm{inp}}, \mathbf{L}_E^1{\downarrow}_{\mathscr{M}^{\mathrm{inp}}}, R_{INP})$ and $\mathbf{L}_2 \in F_{\mathbf{PR}}(AF, \mathscr{M}^{\mathrm{inp}}, \mathbf{L}_E^2{\downarrow}_{\mathscr{M}^{\mathrm{inp}}}, R_{INP})$, it holds that $\mathbf{L}_1 \sqsubseteq \mathbf{L}_2$.

Lemma 8. Let $\mathscr{M}$ be an argumentation multipole (AF, R_{INP}, R_{OUTP}) w.r.t. a set E which is **CO**- univocal under a set of labellings $\mathfrak{L}' \subseteq \mathfrak{L}_E$. Then $\mathscr{M}$ is internally **PR**-homogeneous under $\mathfrak{L}'$.

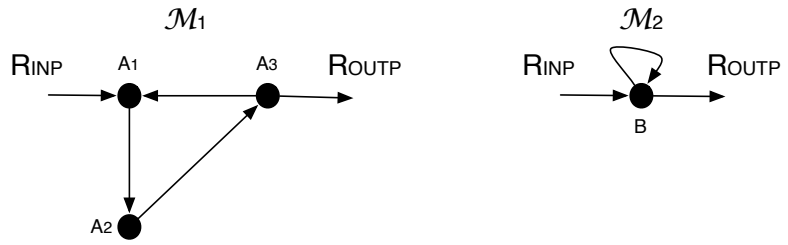


Figure 2.16: Two multipoles that can be safely interchanged under ideal semantics (Example 20).

On this basis, Theorem 15 shows that contextually **CO**-legitimate replacements are **ID**-safe if they involve **CO**-univocal multipoles. Note that the theorem requires the involved multipoles to be **CO**-equivalent under $\mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}}$. In the light of Lemma 6, this is tantamount to requiring them to be **S**-equivalent for any $\mathbf{S} \in \{\mathbf{GR}, \mathbf{PR}, \mathbf{ID}, \mathbf{SST}\}$. We cannot, however, replace $\mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}}$ with e.g. $\mathfrak{L}^{\mathbf{PR}}_{\mathscr{R}}$, since $\mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}}$ may be a strict superset of $\mathfrak{L}^{\mathbf{PR}}_{\mathscr{R}}$.

Theorem 15. *Any contextually* **CO***-legitimate replacement $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$ with invariant set E, such that $\mathscr{M}_1$ and $\mathscr{M}_2$ are* **CO***-univocal under $\mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}}$, is* **ID***-safe.*

As shown in Section 2.9, the previous theorem applies in particular to acyclic argumentation multipoles, while the next example shows that there are cases of equivalent multipoles containing cycles that can be safely interchanged under ideal semantics.

Example 20. It is easy to see that the multipoles $\mathscr{M}_1$ and $\mathscr{M}_2$ shown in Figure 2.16 are **CO**-equivalent and both of them are **CO**-univocal under any set. Thus, by Theorem 15 they can be safely replaced each other under the ideal semantics, i.e. the replacement maintains the labels assigned by the ideal labelling to the arguments of the invariant set. It is also easy to see that the same holds by replacing the three-length cycles in $\mathscr{M}_1$ with any odd-length cycle.

Semi-stable semantics

As in the case of ideal semantics, semi-stable semantics inherits from its lack of decomposability properties the inability of guaranteeing safeness of legitimate replacements: the following example, inspired by Examples 8 and 9, shows that semi-stable semantics is not transparent even w.r.t. $\mathscr{F}_{\mathsf{SCC}}$.

Example 21. Consider the argumentation frameworks AF_1 and AF_2 shown in Figure 2.17, where $AF_2 = T(\mathscr{R})$ with $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$, and the invariant set of the replacement $\mathscr{R}$ is $\{E_1\}$. It is easy to see that $\mathscr{M}_1$ and $\mathscr{M}_2$ are **SST**-equivalent, since $F_{\mathbf{SST}}$ prescribes for $\mathscr{M}_1$ the unique labelling $\{(A_1, \mathtt{out}), (A_2, \mathtt{in})\}$ and for $\mathscr{M}_2$ the unique labelling $\{(B_1, \mathtt{out}), (B_2, \mathtt{in}), (B_3, \mathtt{out})\}$, thus the effect on $\{E_1\}$ is $\{\{(E_1, \mathtt{in})\}\}$ in both cases. As a consequence, the replacement $\mathscr{R}$ is **SST**-legitimate, and it is also easy to see that it is $\mathscr{F}_{\mathsf{SCC}}$-preserving. However, $\mathscr{R}$ is not **SST**-safe, since in AF_1 there is only one semi-stable labelling, namely $\{(A_1, \mathtt{out}), (A_2, \mathtt{in}), (E_1, \mathtt{undec})\}$, which assigns to E_1 the label `undec`, while there are two semi-stable labellings in AF_2, namely $\{(B_1, \mathtt{in}), (B_2, \mathtt{out}), (B_3, \mathtt{undec}), (E_1, \mathtt{out})\}$ and $\{(B_1, \mathtt{out}), (B_2, \mathtt{in}), (B_3, \mathtt{out}), (E_1, \mathtt{undec})\}$, which assign to E_1 the label `out` and `undec`, respectively.

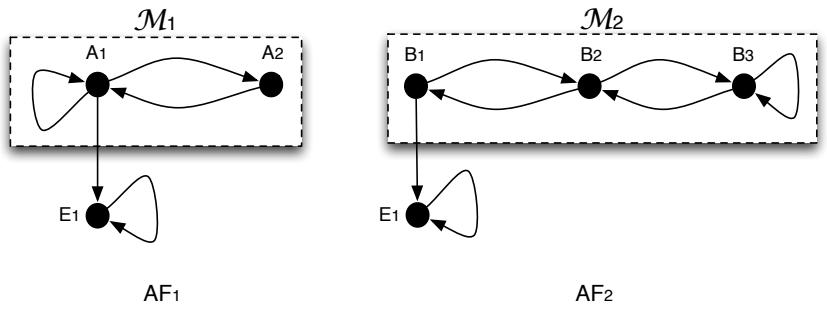


Figure 2.17: Semi-stable semantics is not transparent w.r.t. $\mathscr{F}_{\mathtt{SCC}}$ (Example 21).

The case of acyclic multipoles

It is well-known that an argumentation framework with an acyclic attack relation admits a unique complete labelling which is thus also grounded, preferred, ideal, stable and semi-stable. It is then interesting to specifically consider acyclic multipoles, and to investigate whether they benefit of specific properties as far as replaceability is concerned.

Definition 39. A multipole $\mathscr{M} = (AF, R_{INP}, R_{OUTP})$, where $AF = (Ar, \rightarrow)$, is *acyclic* if there is no sequence $A_1, \ldots, A_n$ of distinct arguments with $A_i \in Ar$ such that $n > 1$, $(A_i, A_{i+1}) \in \rightarrow$ for $1 \leq i < n$, and $(A_n, A_1) \in \rightarrow$.

Note that this definition does not prevent an acyclic multipole to contain self-attacking arguments, i.e. arguments attacking themselves.

The following proposition shows that the property of acyclic frameworks mentioned above can be extended to acyclic multipoles.

Proposition 14. An acyclic argumentation multipole $\mathscr{M} = (AF, R_{INP}, R_{OUTP})$ w.r.t. a set E is **CO**-univocal under any set of labellings $\mathfrak{L}' \subseteq \mathfrak{L}_E$.

The above result entails that all semantics considered in this chapter, with the exception of semi-stable semantics, become strongly transparent in case replacements involve acyclic multipoles. Since admissible, complete, stable and grounded semantics are strongly transparent, it suffices to consider preferred and ideal semantics.

Proposition 15. Any contextually **PR**-legitimate (**ID**-legitimate) replacement $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$ with invariant set E, such that $\mathscr{M}_1$ and $\mathscr{M}_2$ are acyclic, is **PR**-safe (**ID**-safe).

The following example shows that this result cannot be extended to semi-stable semantics, i.e. there are acyclic **SST**-equivalent multipoles that cannot be safely interchanged.

Example 22. Consider the argumentation frameworks AF_1 and AF_2 shown in Figure 2.18, where $AF_2 = T(\mathscr{R})$ with $\mathscr{R} = (AF_1, \mathscr{M}_1, \mathscr{M}_2)$, and the invariant set of $\mathscr{R}$ is $\{E_1, E_2, E_3, E_4\}$. The acyclic multipoles $\mathscr{M}_1$ and $\mathscr{M}_2$ are trivially **SST**-equivalent, since they do not attack E (for both of them, the effect on E includes a unique labelling which assigns to all arguments the label `in`). However, the replacement $\mathscr{R}$ is not **SST**-safe, since there is a unique semi-stable labelling in AF_1, namely $\{(E_1, \mathtt{undec}), (E_2, \mathtt{out}), (E_3, \mathtt{in}), (E_4, \mathtt{out}),$

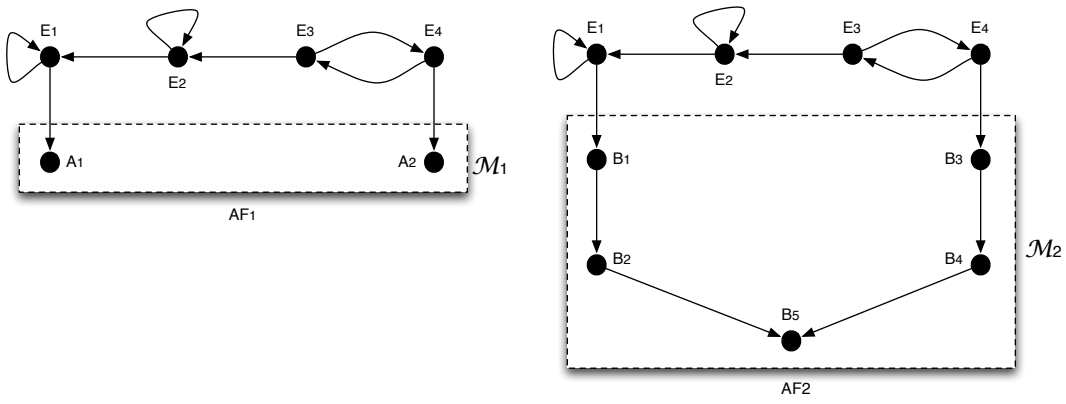


Figure 2.18: Semi-stable semantics is not transparent even considering acyclic multipoles (Example 22).

$(A_1, \texttt{undec}), (A_2, \texttt{in})\}$, while AF_2 admits $\{(E_1, \texttt{undec}), (E_2, \texttt{out}), (E_3, \texttt{in}), (E_4, \texttt{out}), (B_1, \texttt{undec}), (B_2, \texttt{undec}), (B_3, \texttt{in}), (B_4, \texttt{out}), (B_5, \texttt{undec})\}$ and $\{(E_1, \texttt{undec}), (E_2, \texttt{undec}), (E_3, \texttt{out}), (E_4, \texttt{in}), (B_1, \texttt{undec}), (B_2, \texttt{undec}), (B_3, \texttt{out}), (B_4, \texttt{in}), (B_5, \texttt{out})\}$ as the two semi-stable labellings. For instance, argument E_4 is assigned the unique label `out` in AF_1 and the labels `in` and `out` in AF_2.

2.10 Putting modularity at work

As modularity is a very useful and pervasive property, the notions and results introduced in this chapter have an open-ended range of applications. In fact, they can be exploited in all contexts, either theoretical or practical, where a non-monolithic approach is appropriate, ranging from the management of dynamics in argumentation to the study of efficient divide-and-conquer algorithms. While an extensive discussion of related works with pointers to future research directions is given in Section 2.11, in this section we use, as sample case-studies, the tasks of summarization and translation of argumentation frameworks and develop in detail some relevant application examples.

Summarizing argumentation frameworks

In this subsection we illustrate an example of application of the notion of equivalence between argumentation multipoles for the purpose of summarization of argumentation frameworks. In particular we take from the literature two argument-based recontructions of the court's decision of the Popov v. Hayashi case and show that, in spite of many differences in the details, they can be reduced to a comparable basic structure through considerations based on multipole equivalence.

We borrow a synthetic description of the facts originating the case from [324]. "The case concerned the possession of the baseball which Barry Bonds hit for his record breaking 73rd home run in the 2001 season. Such a ball is very valuable (Mark McGwire's 1998 70th home run ball sold at auction for $3,000,000). When the ball was struck into the crowd, Popov caught it in the upper part of the webbing of his baseball glove. Such a catch, a snowcone catch because the ball is not fully in the mitt, does not give certainty of retaining control of the ball, particularly since Popov was stretching and may have fallen. However, Popov

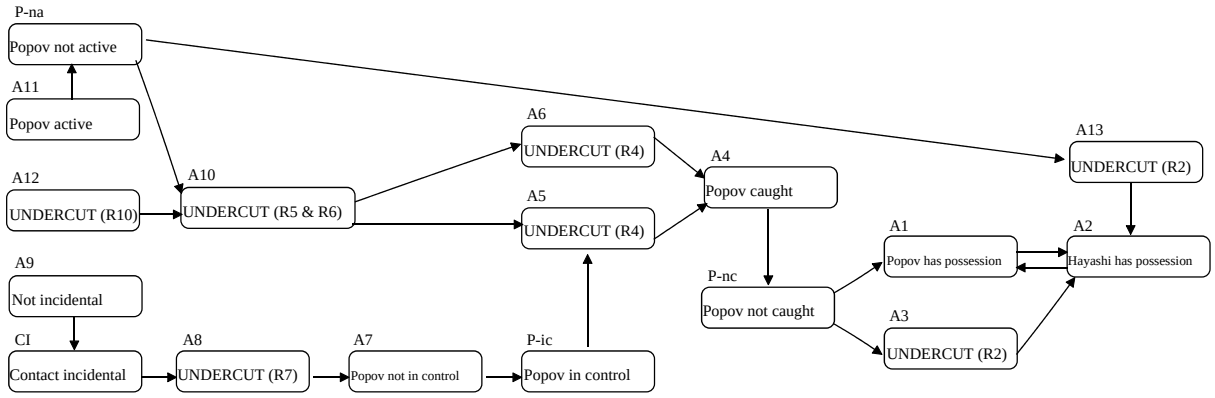


Figure 2.19: The argumentation framework AF_J for the Popov v. Hayashi case from [324].

was not given the chance to complete his catch since, as it entered his glove, he was tackled and thrown to the ground by others trying to secure the ball, which became dislodged from his glove. Hayashi (himself innocent of the attack on Popov), then picked up the ball and put it in his pocket, so securing possession."

Popov then claimed possession of the ball and sued Hayashi. The court finally decided that the ball should be sold and the proceeds divided between the two.

The rather articulated motivations underlying the decision have attracted the attention of researchers and have been the subject of several chapters, culminating in a special issue of the *Artificial Intelligenge and Law* journal devoted to the modelling of this case [11]. In the following subsections we present the argument-based formalizations provided by Wyner and Bench-Capon [324] and by Prakken [262] respectively. Then we show how the notions and results presented in previous sections can be used to summarize the two formalizations and simplify their comparison.

For the sake of uniformity with the original formalizations, in the following we will sometimes refer to the extension-based rather than the labelling-based approach. In particular, an **S** extension (e.g. the grounded extension) is the set of arguments labelled `in` by an **S** labelling (e.g. the grounded labelling).

The formalization by Wyner and Bench-Capon

In [324] the legal analysis of the case is synthesized by the argumentation framework presented in Figure 2.19 (the chapter also presents an analysis of the values underlying the final decision using the formalism of value-based argumentation frameworks, which is beyond the scope of the present chapter). In the original figure of [324] the boxes representing arguments are labeled with an identifier Ax, where x is a number, while a few other boxes have no label and contain a statement corresponding to the conclusion of the argument. In Figure 2.19 all arguments have both a label (on top of the box and corresponding to the original one where present) and a text synthesizing their conclusion. Each argument labeled as Ax derives from the application of a rule with some premises and a conclusion, while the other four arguments are intended to represent default answers to some questions: quoting [324], "if the argument is not defeated, the contrary has not been shown". The conclusion of an argument may correspond to the undercut of some rule. An argument attacks another argument if the conclusion of the former contradicts the conclusion or some premise of the

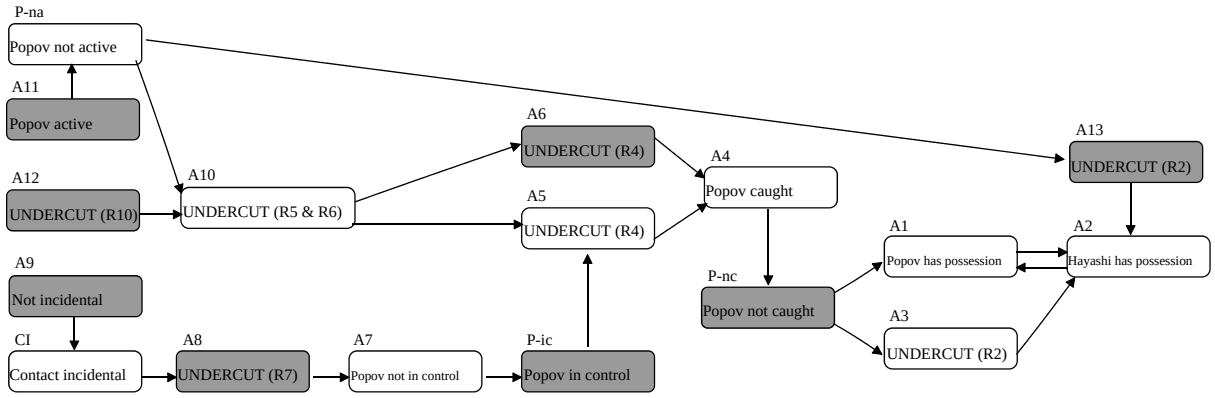


Figure 2.20: The only extension of AF_J.

latter or undercuts the rule used for its construction. Default arguments can only attack another argument on its premises. Turning to a quick explanation of Figure 2.19, we can proceed backwards starting from the mutually attacking arguments A1 and A2, concerning who has possession of the ball. A2 is undercut by A13: the rule that Hayashi has possession of the ball because he retrieved it is not applicable given that Popov was active in catching the ball before Hayashi retrieved it. A13 is attacked by the default argument P-na, which is in turn attacked by A11 based on factual evidence of the snowcone catch. A2 is also undercut by A3, whose premise (by the way, the same as of A1) is that Popov caught the ball before Hayashi. However both A1 and A3 are attacked by the default argument that the ball was not caught by Popov. This is in turn attacked by A4, based on the fact that the ball was in Popov's glove. A4 is undercut by A5 and A6, the former based on the fact that the ball was still in motion, the latter on the fact that Popov was not in control of the ball. Both A5 and A6 are undercut by A10 based on the fact that Popov was active. A10 is hence attacked by the default argument P-na and is also undercut by A12, based on the custom and practice of the stands in baseball. Moreover A5 is attacked by the default argument P-ic, which is attacked by A7 based on the fact that Popov did not retain the ball in the glove. A7 is undercut by A8, based on the fact that Popov lost the ball due to an intentional contact of other people. Finally, A8 is attacked by the default argument CI which is in turn attacked by A9 based on factual evidence that Popov was assaulted.

It can be seen that for the argumentation framework represented in Figure 2.19 the grounded extension is also the only complete, stable, semi-stable, ideal and preferred extension. It consists of the arguments A9, A11, A12, A13, A6, A8, P-ic, P-nc, which are evidenced in grey in Figure 2.20. We note that both A1 and A2 are rejected according to any semantics, leaving the issue of the possession of the ball unresolved.

The formalization by Prakken

The reconstruction of the case given in [262] adopts ASPIC+, which is essentially a rule-based formalism for the construction of arguments and the identification of their subargument and attack relations. It is worth remarking that the latter takes into account the former: if an argument attacks another argument then it attacks also all its superarguments. In ASPIC+ argument status evaluation follows Dung's approach: an argumentation framework consisting only of the arguments and their attack relations can be derived and

then the semantics deemed most appropriate can be applied.

Coming back to Popov and Hayashi, the reconstruction of [262] covers a lot of details concerning argument construction and, as such, is much more articulated than the one of [324] as shown by Figures 2.21 and 2.22 which correspond to the aggregation of five distinct but linked figures included in [262]. Direct subargument relationships are represented by dashed lines ending with a solid dot on the superargument, attack relationships are represented by solid arrows ending on the attacked argument. The text in an argument box essentially gives an idea on its conclusion. Figure 2.21 is referred to as the upper part, while Figure 2.22 is referred to as the lower part, they are linked only by two subargument relations: VR-MC8 and VR-r1 in Figure 2.22 are direct subarguments respectively of EQ and H-hr in Figure 2.21.

For a detailed description of the whole reconstruction, which is clearly beyond the scope of the present chapter, the reader is referred to [262]. At a general level we can observe that:

- a lot of attention is reserved to issues concerning the validity of rules (sometimes based in turn on the validity of other rules), their adoption and their applicability to the case into question;
- the lower part (Figure 2.22) essentially concerns the question whether Popov gained possession of the ball. There are two alternative reasoning lines leading to this conclusion, composed respectively by arguments VR-cs4, P-cc(1), P-ca(1), P-ph(1), P-hp(1), and P-wit, P-cb, VR-cs2, P-cc(2), P-ca(2), P-ph(2), P-hp(2). Both lines are defeated, the former by argument NV-cs4 stating the invalidity of the rule cs4 which is the starting point of the whole line, the latter by argument P-inc stating that Popov's testimony, on which the whole line is based, is not credible.
- the upper part (Figure 2.21) essentially concerns the action to be taken: three mutually exclusive alternatives (corresponding to the three mutually attacking arguments H-hr, H-nr, and EQ) are considered: Hayashi has to return the ball, Hayashi has not to return the ball, the ball is equally shared. Each of the three arguments is derived through a quite articulated reasoning line. Both H-hr and H-nr are defeated, the former by argument NV-rp, stating that the rule rp is not valid, the latter by argument NA-r4, stating that the rule r4 is not applicable.

If one considers the attack relations only (i.e. focuses on the argumentation framework to be used for argument status evaluation) the picture is simplified, as shown in Figure 2.23, since a large number of arguments are neither attacking nor attacked by others. It can be seen that for the argumentation framework represented in Figure 2.23 the grounded extension is also the only complete, stable, semi-stable, ideal and preferred extension and consists of the arguments evidenced in grey. We note that of the three arguments corresponding to the possible final decisions both H-hr and H-nr are rejected, while EQ is accepted.

Summarizing and comparing the two formalizations

We can now use considerations based on the equivalence properties examined in the previous sections to identify some fundamental similarities between the two reconstructions of the case.

As to the argumentation framework $AF_J = (Ar, \rightarrow)$ of Figure 2.19, let us start by considering the argumentation multipole $\mathscr{M}_1 = (AF_J{\downarrow}_{\{\text{A11, P-na}\}}, \emptyset, \{(\text{P-na, A10}), (\text{P-na, A13})\})$ with respect to $E_1 = Ar \setminus \{\text{A11, P-na}\}$. It is rather easy to see that for any labeling $\mathbf{L}_{E_1}$ of E_1 (actually irrelevant since the multipole does not receive attacks) and for any semantics $\mathbf{S}$ (all behave the same on such a simple subframework) it holds that $\mathbf{S}\text{-}\mathtt{eff}_{E_1}(\mathscr{M}_1, \mathbf{L}_{E_1}) = \{(\text{A}, \mathtt{in}) \mid A \in E_1\} = \mathbf{S}\text{-}\mathtt{eff}_{E_1}(\mathscr{M}_\emptyset, \mathbf{L}_{E_1})$.

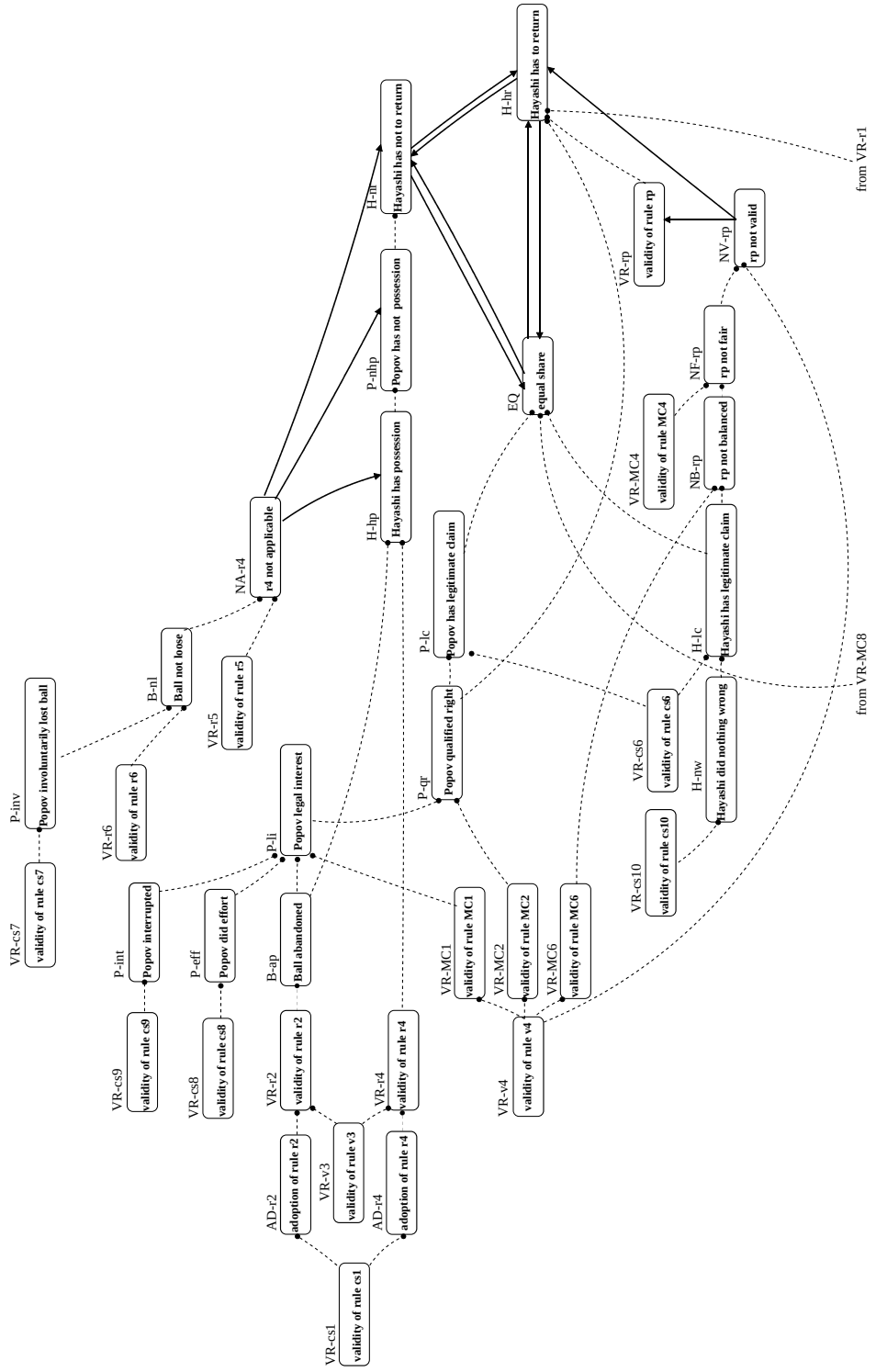


Figure 2.21: Upper part of the representation of the Popov v. Hayashi case from [262].

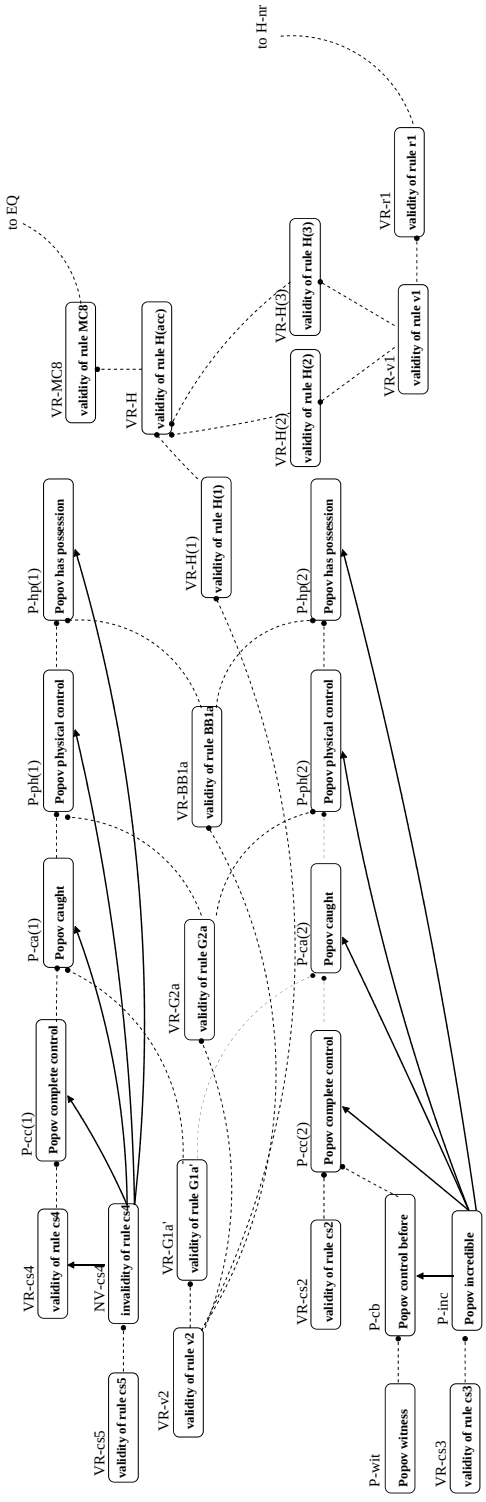


Figure 2.22: Lower part of the representation of the Popov v. Hayashi case from [262].

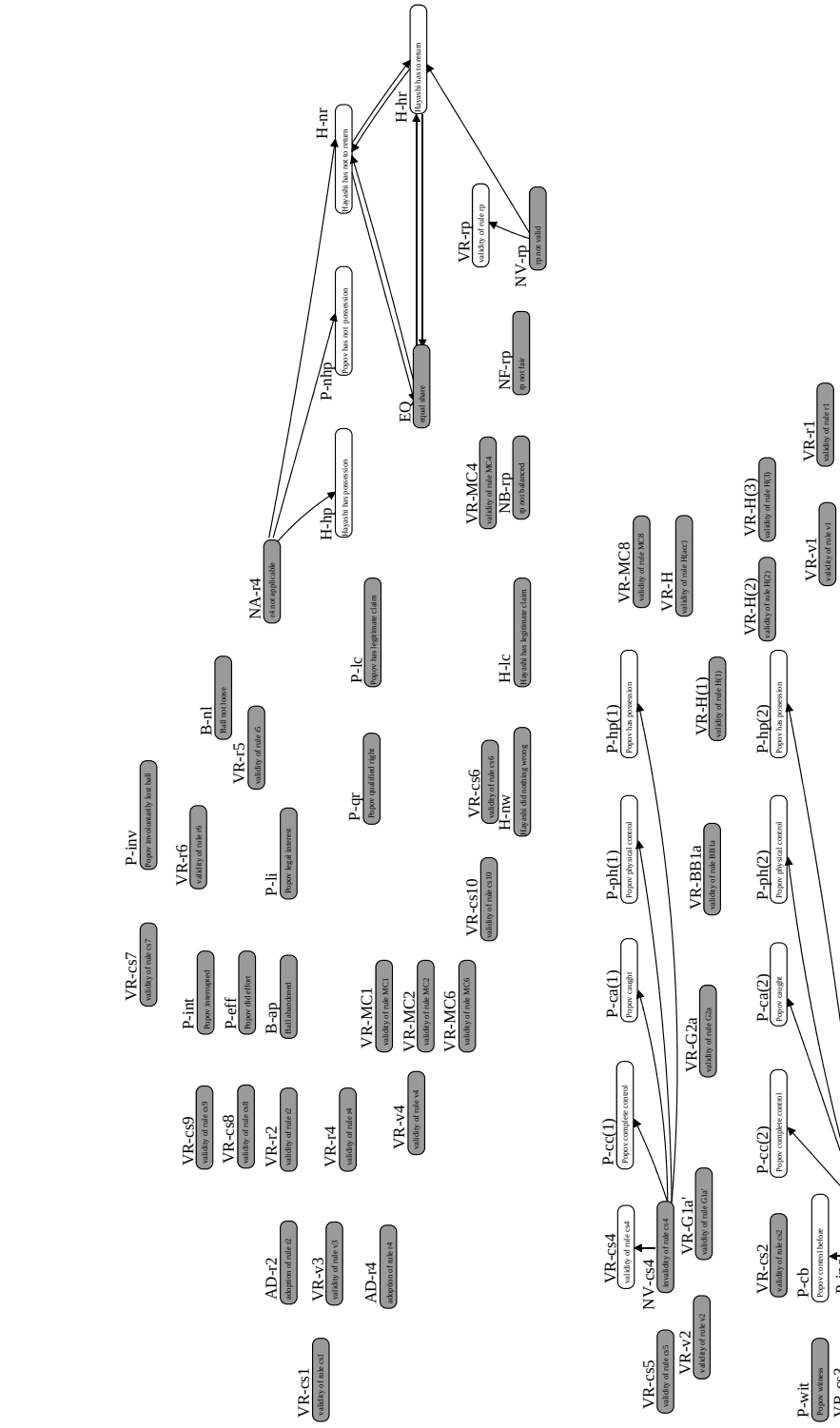


Figure 2.23: The representation of the Popov v. Hayashi case from [262] without the subargument relation.

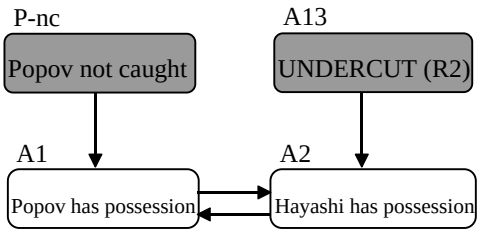


Figure 2.24: The argumentation framework AF_J^- summarizing the reconstruction from [324].

In other words, the multipole $\mathscr{M}_1$ is **S**-equivalent to the empty multipole for any semantics **S**. It follows that the replacement $\mathscr{R} = (AF_J, \mathscr{M}_1, \mathscr{M}_\emptyset)$ is **S**-legitimate. Intuitively this means that the arguments $A11$ and $P-na$ can be canceled from AF_J without any consequence on the evaluation of other arguments, provided that a suitable transparency property holds for **S**. Since both multipoles $\mathscr{M}_1$ and $\mathscr{M}_\emptyset$ are acyclic, the results summarized in Table 2.2 ensure that the replacement is safe for any semantics considered in this chapter except semi-stable semantics (by the way, the replacement is safe also for semi-stable semantics, given that in this case its labellings coincide with stable labellings).

Iterating the same kind of reasoning, it can be seen that the following pairs of arguments can progressively (and safely) be cancelled: {A12, A10}, {A9, CI}, {A8, A7}, {P-ic, A5}, {A6, A4}. In virtue of Proposition 13 we have that we can safely restrict AF_J to the set of arguments $E^* = \{$A1, A2, A3, A13, P-nc$\}$ without affecting the labellings of the arguments in E^*. This could have been done (in a single, more laborious, step) also showing that the big multipole consisting of the set of arguments {A11, P-na, A12, A10, A9, CI, A8, A7, P-ic, A5, A6, A4} is **S**-equivalent to the empty multipole.

Assuming that the main focus concerns the evaluation of arguments A1 and A2, we can also see that A3 can be suppressed in $AF_J^* = AF_J{\downarrow_{E^*}}$: given the multipole $\mathscr{M}_2 = (AF_J^*{\downarrow_{\{\text{A3, P-nc}\}}}, \emptyset, \{(\text{P-nc, A1}), (\text{A3, A2})\})$ with respect to $E_2 = \{$A1, A2, A13$\}$, it is again easy to see that for any (actually irrelevant) labelling $\mathbf{L}_{E_2}$ of E_2 and for any semantics **S** it holds that $\mathbf{S}\text{-}\mathtt{eff}_{E_2}(\mathscr{M}_2, \mathbf{L}_{E_2}) = \{(\text{A1}, \mathtt{out}), (\text{A2}, \mathtt{in}), (\text{A13}, \mathtt{in})\} = \mathbf{S}\text{-}\mathtt{eff}_{E_2}(\mathscr{M}_3, \mathbf{L}_{E_2})$ where $\mathscr{M}_3 \triangleq ((\{\text{P-nc}\}, \emptyset), \emptyset, \{(\text{P-nc, A1})\})$. Using again the fact that both $\mathscr{M}_2$ and $\mathscr{M}_3$ are acyclic we get that the replacement is safe, i.e. that A3 can be cancelled.

Summing up, we get the simplified argumentation framework AF_J^- shown in Figure 2.24 which, for any semantics considered in this chapter, is equivalent to the original one as far as the evaluation of the remaining arguments is concerned.

Turning now to the argumentation framework AF_K of Figure 2.23, we first note that all the isolated (i.e. both unattacking and unattacked) arguments can be suppressed. This follows from the fact that, for any semantics **S** and for any argumentation framework AF_U such that $\mathbf{L}_\mathbf{S}(AF_U) \neq \emptyset$, given the multipole $\mathscr{M}_U = (AF_U, \emptyset, \emptyset)$ with respect to any (actually irrelevant) set E, for any labeling $\mathbf{L}_E$ of E it holds that $\mathbf{S}\text{-}\mathtt{eff}_E(\mathscr{M}_U, \mathbf{L}_E) = \{(A, \mathtt{in}) | A \in E\} = \mathbf{S}\text{-}\mathtt{eff}_{E_1}(\mathscr{M}_\emptyset, \mathbf{L}_E)$.

Supposing that the main interest concerns the final decision, i.e. the evaluation of the arguments H-hr, H-

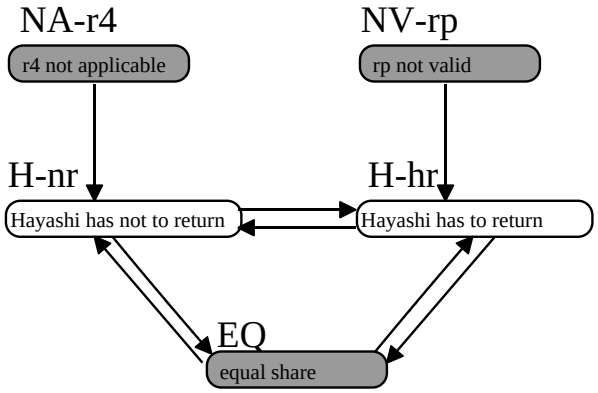


Figure 2.25: The argumentation framework AF_K^- summarizing the reconstruction from [262].

nr and EQ, and using the same reasoning as above we can also see that all the arguments concerning the issue of Popov's possession, not attacking nor being attacked by arguments outside the set, can be suppressed.

Then, using a reasoning which is completely analogous to the one applied to the multipole $\mathscr{M}_2$ above, we can also suppress the arguments H-hp, P-nhp, and Vr-rp, getting finally the argumentation framework AF_K^- represented in Figure 2.25.

Comparing now Figures 2.24 and 2.25 we observe that:

- arguments A1 and A2 in AF_J^- correspond respectively to arguments H-hr and H-nr in AF_K^- and have the same status of rejected;
- similarly, we can also say that arguments P-nc and A13 in AF_J^- correspond respectively to arguments NV-rp and NA-r4 in in AF_K^-;
- the argument EQ of AF_K^- has no counterpart in AF_J^- due to the fact that in [324] the final decision is represented only in the context of the value-based formalization.

Leaving apart EQ, we note therefore a basic structural similarity between the two simplified frameworks: in both reconstructions the arguments corresponding to giving the ball to one of the contendants are rejected due to one main reason. One may then wonder whether the reasons for these rejections are actually the same in the two reconstructions.

As to the rejection of the decision in favor of Hayashi, in AF_J^- it is due to the undercut of A2 by A13, which is based on the fact that Popov was "ably and actively engaged in establishing control" of the ball. Similarly, in AF_K^- the rejection of H-nr is due to the fact that a rule used to derive that Hayashi has possession of the ball, is shown not to be applicable in this case through argument NA-r4, based on the fact that the ball was not loose (due to the previous attempt of Popov) when Hayashi retrieved it.

While basically similar as far as the previous point is concerned, the two reconstructions turn out to be different as to the rejection of the decision in favor of Popov: in AF_J^- A1 is attacked by P-nc which

corresponds to the conclusion that Popov did not catch the ball, thus denying the premise of A1, while in AF_K^- H-hr is attacked by NV-rp, which concerns the validity of the rule rp. It is interesting to note that in [262] the argument NV-rp is essentially based on the fact that "rule rp does not promote fundamental fairness as regards Popov's claim" and that, indeed, fairness is the primary value considered in the value-based part of [324] as a justification of the final decision.

Thus the difference arises from the fact that in the formalism adopted in [262] reasoning about values is embedded into arguments that are at the same level as other arguments, while in [324] reasoning about values is carried out in a separate layer. A discussion about the pros and cons of either approach to deal with values is clearly out of the scope of this chapter.

To conclude this section we remark that the identification of some basic commonalities and differences between two argument-based reconstructions of a real law case has been greatly simplified by the possibility to summarize frameworks in a general and technically sound way. In this perspective the notion of argumentation multipole and the decomposability and equivalence properties investigated in this chapter can be regarded as enabling techniques for the investigation of methods for (possibly automated) analysis, synthesis and comparison of argumentation frameworks.

Translations of argumentation frameworks

Translating an argumentation framework AF_1 into another framework AF_2 such that AF_2 has some desirable features and, at the same time, preserves some specific properties of AF_1 is a generic problem with significant theoretical and practical implications. In particular in [140] the problem of *intertraslatability* is considered, which is defined as follows: "Given an argumentation framework F and argumentation semantics σ and σ', find a function Tr such that the σ-extensions of F are in certain correspondence to the σ'-extensions of $Tr(F)$." As a matter of fact, in [140] *modularity* is one of the general requirements of a translation procedure, informally stated as "the translation can be done independently for certain parts of the framework". While this generic notion may have different technical counterparts depending on the kind of translation addressed, our results provide a systematic and sound basis for ensuring modularity in any context where there is an interest in replacing a subframework with a translated counterpart. A broad investigation of this issue is clearly a matter for future work, here we provide two specific examples taken from the literature: the former concerns a subframework replacement considered in the context of the analysis of the properties of weighted argument systems, while the latter concerns the translation (also called *flattening*) of argumentation frameworks with attacks to attacks into "traditional" Dung's frameworks.

Reducing the attacks involving single arguments under grounded semantics

A weighted argument system (WAS in the following), as defined in [134], is basically an argumentation framework with a numerical weight (actually a non-negative real number) attached to each attack. In the analysis of the computational properties of WASs, it turns out to be convenient to consider a translation from a WAS into another one such that no argument attacks or is attacked by more than 2 arguments and some conditions are satisfied. Leaving apart the aspects of the translation and the conditions involving weights, which are not relevant to the present chapter, basically the translation described in [134] involves replacing the subframework consisting of an argument z receiving more than two attacks (from arguments $y_1, \ldots, y_k$) with a subframework with additional arguments p_1 and q_1 where z receives only two attacks (from p_1 and and y_1), while p_1 is attacked by q_1 and q_1 is attacked by the arguments $y_2, \ldots, y_k$ (see Figure 2.26). The replacement can then be iterated focusing on q_1 and adding p_2 and q_2 until q_{k-2} is only attacked by y_{k-1} and

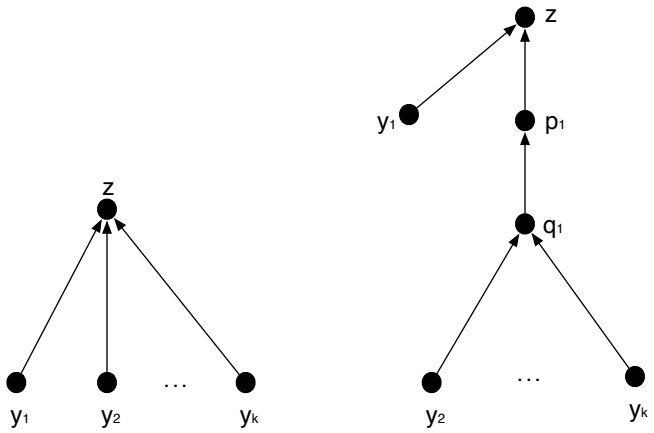


Figure 2.26: A translation in the context of WAS.

y_k. The claim (proved in [134] as part of Lemma 1) is that the grounded extension of the original framework is the same as the grounded extension of the framework resulting from the replacements mentioned above. Note that Lemma 1 of [134] concerns an arbitrary WAS, i.e. its hypotheses do not put any restriction on other attacks present in the original framework. In particular, as explicitly remarked in [134], there can be attacks between some of the attackers of z, but also (not explicitly remarked in [134]) z might counterattack some of its attackers or there could be longer loops involving z, some of its attackers and possibly other arguments in the framework.

Given these remarks, the proof of Lemma 1 provided in [134] is not completely satisfactory: it consists in local considerations on the arguments involved in the replacement described above without dealing with possible effects involving other arguments in the framework. The absence of these effects, however, can not be taken for granted. To give an example, when considering (to contradiction) a generic argument x included in the grounded extension of the original framework but not in the grounded extension of the translated framework it is stated that this implies that there is an attacker u of x in the translated framework such that (u,x) was not an attack in the original framework. This immediately leads to identify x as z and u as p_1 and to apply only local considerations. However, in general, an argument might be excluded from the grounded extension not just because it has an additional attacker but also because one of its attackers has a different justification state in the new framework. In a sense, the proof of Lemma 1 of [134] seems to implicitly assume the property of transparency of grounded semantics (which, of course is not obvious *per se*) and (partially) shows a sort of local equivalence of the original fragment and of its translated counterpart.

Actually, the result of Lemma 1 of [134] is valid and this can be shown in a relatively straightforward way using the results of the present chapter. First, given that grounded semantics is strongly transparent, to obtain the result it is sufficient to show that the translation step depicted in Figure 2.26 involves the replacement of an argumentation multipole with another one which is Input/Output **GR**-equivalent. The fact that the translation may involve several such steps is then covered by the result of Proposition 13.

As to the identification of the equivalent multipoles $\mathscr{M}_1 = (AF^1, R^1_{INP}, R^1_{OUTP})$ and $\mathscr{M}_2 = (AF^2, R^2_{INP}, R^2_{OUTP})$, observe that the basic idea consists in replacing the argument z with the attack chain composed by the three arguments p_1, q_1, and z itself within an arbitrary argumentation framework $AF = (Ar, \rightarrow)$ where $\{y_1, \ldots, y_k\}$ is the set of attackers of z with $k > 2$. Then $\mathscr{M}_1$ and $\mathscr{M}_2$ are defined with respect to the same

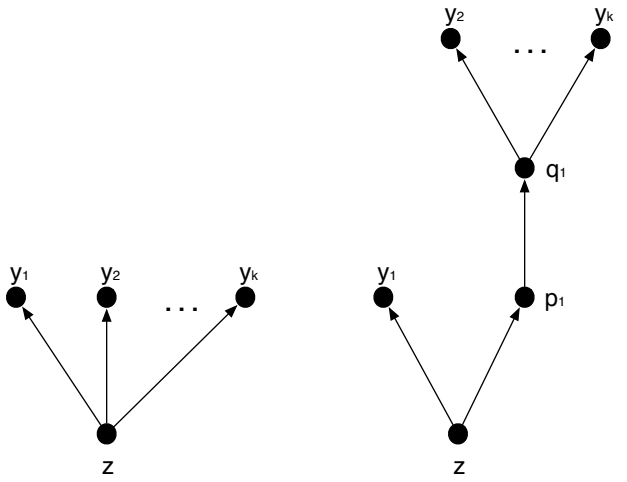


Figure 2.27: Another translation in the context of WAS.

invariant set $E = Ar \setminus \{z\}$, and the relevant frameworks are $AF_1 = (\{z\}, \emptyset)$ and $AF_2 = (\{p_1, q_1, z\}, \{(q_1, p_1), (p_1, z)\})$. Moreover $\mathscr{M}_1$ and $\mathscr{M}_2$ have the same output relation: $R^1_{OUTP} = R^2_{OUTP} = \rightarrow \cap (\{z\} \times Ar)$, while they differ in the input relation: $R^1_{INP} = \{(y_i, z) \mid 1 \leq i \leq k\}$; $R^2_{INP} = \{(y_1, z)\} \cup \{(y_i, q_1) \mid 2 \leq i \leq k\}$. We have now to show that $\mathscr{M}_1$ and $\mathscr{M}_2$ are **GR**-equivalent, i.e. that for any labelling $\mathbf{L}_E \in \mathfrak{L}_E$, $\mathbf{GR}\text{-}\mathtt{eff}_E(\mathscr{M}_1, \mathbf{L}_E) = \mathbf{GR}\text{-}\mathtt{eff}_E(\mathscr{M}_2, \mathbf{L}_E)$, which, recalling Definition 31, amounts to show that $\{\mathtt{eff}_E({\mathscr{M}_1}^{\mathrm{outp}}, \mathbf{L}{\downarrow}_{\mathscr{M}_1 \mathrm{outp}}, R^1_{OUTP}) \mid \mathbf{L} \in F_{\mathbf{GR}}(AF^1, {\mathscr{M}_1}^{\mathrm{inp}}, \mathbf{L}_E{\downarrow}_{\mathscr{M}_1 \mathrm{inp}}, R^1_{INP})\} = \{\mathtt{eff}_E({\mathscr{M}_2}^{\mathrm{outp}}, \mathbf{L}{\downarrow}_{\mathscr{M}_2 \mathrm{outp}}, R^2_{OUTP}) \mid \mathbf{L} \in F_{\mathbf{GR}}(AF^2, {\mathscr{M}_2}^{\mathrm{inp}}, \mathbf{L}_E{\downarrow}_{\mathscr{M}_2 \mathrm{inp}}, R^2_{INP})\}$.

First note that, since ${\mathscr{M}_1}^{\mathrm{outp}} = {\mathscr{M}_2}^{\mathrm{outp}} = \{z\}$ and $R^1_{OUTP} = R^2_{OUTP}$, both $\mathtt{eff}_E({\mathscr{M}_1}^{\mathrm{outp}}, \mathbf{L}{\downarrow}_{\mathscr{M}_1 \mathrm{outp}}, R^1_{OUTP})$ and $\mathtt{eff}_E({\mathscr{M}_2}^{\mathrm{outp}}, \mathbf{L}{\downarrow}_{\mathscr{M}_2 \mathrm{outp}}, R^2_{OUTP})$ are totally determined by the label assigned to z by $F_{\mathbf{GR}}$ given the labelling of the arguments in the input set $\{y_1, \ldots, y_k\}$ (which is the same for both multipoles).

Now it is easy to see that the label assigned to z is the same for any labelling of the arguments $\{y_1, \ldots, y_k\}$ considering three basic cases: i) $\exists y_i \in \{y_1, \ldots, y_k\} : Lab(y_i) = \mathtt{in}$; ii) $\forall y_i \in \{y_1, \ldots, y_k\} : Lab(y_i) = \mathtt{out}$; iii) $\nexists y_i \in \{y_1, \ldots, y_k\} : Lab(y_i) = \mathtt{in} \wedge \exists y_i \in \{y_1, \ldots, y_k\} : Lab(y_i) = \mathtt{undec}$.

In the case i), clearly z is assigned the label `out` by $F_{\mathbf{GR}}$ in $\mathscr{M}_1$ and this also holds in $\mathscr{M}_2$ since either z is attacked directly by an argument labelled `in` (if $Lab(y_1) = \mathtt{in}$) or, if this is not the case, necessarily $\exists y_i \in \{y_2, \ldots, y_k\} : Lab(y_i) = \mathtt{in}$ and then $Lab(q_1) = \mathtt{out}$, $Lab(p_1) = \mathtt{in}$, $Lab(z) = \mathtt{out}$.

In the case ii), clearly z is assigned the label `in` by $F_{\mathbf{GR}}$ in $\mathscr{M}_1$ and this also holds in $\mathscr{M}_2$: given $\forall y_i \in \{y_1, \ldots, y_k\} : Lab(y_i) = \mathtt{out}$ it follows $Lab(q_1) = \mathtt{in}$, $Lab(p_1) = \mathtt{out}$ and then both attackers (y_1 and p_1) of z are labelled `out` and z is labelled `in`.

In the case iii), clearly z is assigned the label `undec` by $F_{\mathbf{GR}}$ in $\mathscr{M}_1$. As to $\mathscr{M}_2$, first note that y_1 is either labelled `undec` or `out` (in the latter case necessarily $\exists y_i \in \{y_2, \ldots, y_k\} : Lab(y_i) = \mathtt{undec}$). Moreover, q_1 is either labelled `in` or `undec` and consequently p_1 is labelled `out` or `undec` (both are necessarily `undec` if $Lab(y_1) = \mathtt{out}$). Summing up, z is either attacked by two arguments labelled `undec` or by one labelled `undec` and one labelled `out` and hence is labelled `undec` by $F_{\mathbf{GR}}$, as required.

A similar reasoning applies to the case where an argument attacks more than two other arguments, using the replacement sketched in Figure 2.27.

Flattening attacks to attacks

In recent years several extensions of Dung's framework encompassing attacks to attacks have been considered, like the *EAF* (*Extended Argumentation Framework*) formalism [228], mainly designed for the purpose of preference modelling, and the more general (as, differently from *EAF*, they allow unlimited recursion of attacks on attacks) *AFRA* (*Argumentation Framework with Recursive Attacks*) [22, 21] and *HLAF* (*Higher Level Argumentation Framework*) [150].

For the sake of keeping the example compact, we focus here on the *EAF* formalism whose definition (taken from [228]) is given below.

Definition 40. An Extended Argumentation Framework (EAF) is a tuple $(Args, \mathscr{R}, \mathscr{D})$ such that $Args$ is a set of arguments and:

- $\mathscr{R} \subseteq Args \times Args$
- $\mathscr{D} \subseteq Args \times \mathscr{R}$
- if $(X,(Y,Z)),(X',(Z,Y)) \in \mathscr{D}$ then $(X,X'),(X',X) \in \mathscr{R}$.

As typical in any kind of extension of Dung's framework, there is an interest in defining a translation procedure from the extended formalism to the basic one. This is useful for several purposes, including the opportunity to reuse or adapt, in the extended context, the large corpus of theoretical results available in Dung's framework, in particular as far as computational complexity is concerned.

In the case of attacks to attacks, as to our knowledge, two main translation procedures have been proposed in the literature. The first procedure (considered with some slight variants in [229, 150, 54]) involves replacing an attacked attack with an attack chain consisting in two additional arguments, then every attack towards the replaced attack becomes an attack towards the second additional argument. The second procedure (considered in [22, 21]) involves replacing an attack with a single new argument, with a proper rearrangement of the incoming and outcoming attacks involving it. We present in the following the formal definition of these procedures tailored to the case of EAF.

Definition 41. Let $\Gamma = (Args, \mathscr{R}, \mathscr{D})$ be an EAF and let us define $\mathscr{D}^{\rightarrow}(\Gamma) = \{(A,B) | \mathscr{D} \cap (Args \times \{(A,B)\}) \neq \emptyset\}$ i.e. the set of attacks receiving at least an attack according to the relation $\mathscr{D}$.

- The *chain-style* flattening of Γ is the argumentation framework $AF_c^{\Gamma} = (Args_c, \rightarrow_c)$ where $Args_c = Args \cup \{X_{A,B}, Y_{A,B} | (A,B) \in \mathscr{D}^{\rightarrow}(\Gamma)\}$ and $\rightarrow_c = \mathscr{R} \cup \{(A,X_{A,B}),(X_{A,B},Y_{A,B}),(Y_{A,B},B) | (A,B) \in \mathscr{D}^{\rightarrow}(\Gamma)\} \cup \{(C,Y_{A,B}) | (C,(A,B)) \in \mathscr{D}\}$
- The *single-argument* flattening of Γ is the argumentation framework $AF_{sa}^{\Gamma} = (Args_{sa}, \rightarrow_{sa})$ where $Args_{sa} = Args \cup \{\overline{AB} | (A,B) \in \mathscr{D}^{\rightarrow}(\Gamma)\}$ and $\rightarrow_{sa} = (\mathscr{R} \setminus \{(A,B) | (A,B) \in \mathscr{D}^{\rightarrow}(\Gamma)\}) \cup \{(\overline{AB},B) | (A,B) \in \mathscr{D}^{\rightarrow}(\Gamma)\} \cup \{(D,\overline{AB}) | (A,B) \in \mathscr{D}^{\rightarrow}(\Gamma) \wedge (D,A) \in \mathscr{R}\} \cup \{(C,\overline{AB}) | (C,(A,B)) \in \mathscr{D}\}$.

In words, in chain-style flattening two arguments $X_{A,B}$ and $Y_{A,B}$ are added in replacement of every attacked attack (A,B) (with $A, X_{A,B}, Y_{A,B}, B$ forming an attack chain) and the arguments attacking (A,B) according to the relation $\mathscr{D}$ of Γ attack $Y_{A,B}$ in AF_c (while the attacks between arguments in $\mathscr{R}$ remain the

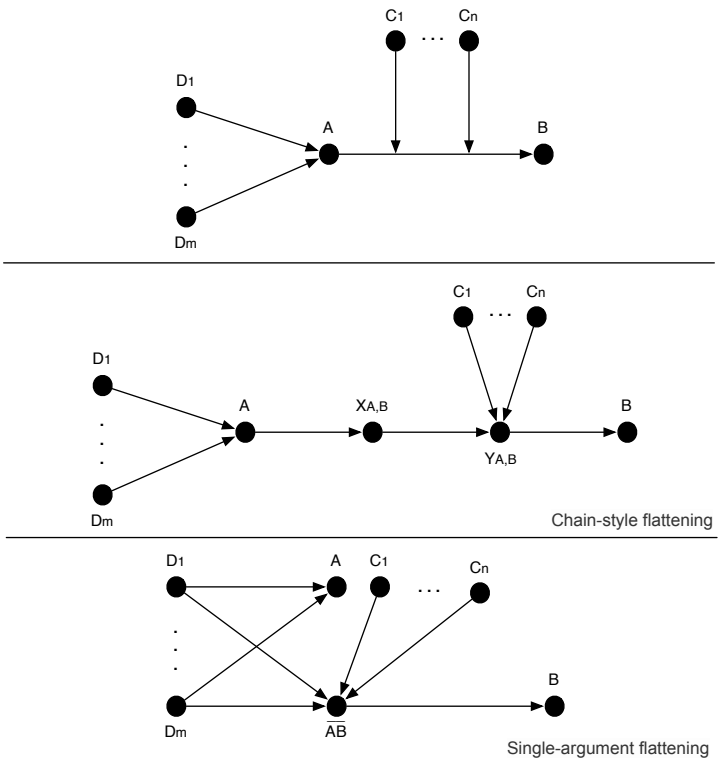


Figure 2.28: The two translation procedures of attacks to attacks.

same). In single-argument flattening every attacked attack (A,B) is replaced by a single argument $\overline{AB}$ which attacks B (instead of A) and is attacked by all attackers of A in $\mathscr{R}$ and by all attackers of (A,B) in $\mathscr{D}$.

The two translation procedures are illustrated in Figure 2.28.

Of course one may wonder whether the operational differences in the two flattening procedures give rise to any actual difference in the final outcome (i.e. in the justification status of the arguments originally included in Γ) or, indeed, the two flattened frameworks treat the arguments originally included in Γ in the same way, showing that the two procedures, different as they are, basically capture the same intuition.

To answer this question first observe that, given an EAF $\Gamma = (Args, \mathscr{R}, \mathscr{D})$, both the argumentation frameworks AF_c^Γ and AF_{sa}^Γ include all the original arguments $Args$ and that they locally differ in correspondence of the additional arguments used to represent the elements of $\mathscr{D}^{\rightarrow}(\Gamma)$. So AF_{sa}^Γ can be obtained from AF_c^Γ (and vice versa) through the replacements of a (possibly quite articulated) multipole $\mathscr{M}_c$ with another multipole $\mathscr{M}_{sa}$, both referring to the same set of invariant arguments $Args$. Thus, answering the question amounts to analysing the safeness of this replacement, based in turn on the equivalence between these multipoles.

First, observe that the multipoles $\mathscr{M}_c$ and $\mathscr{M}_{sa}$ consist of the union of $|\mathscr{D}^{\rightarrow}(\Gamma)|$ disjoint and non-interacting "submultipoles" each having the form illustrated in Figure 2.28. In virtue of Proposition 13, we can then consider a sequence of (similar) replacements leading from AF_{sa}^Γ to AF_c^Γ and, to ensure that the whole sequence of replacements is safe (as far as the arguments $Args$ are concerned), it is sufficient to show that each single step is safe, i.e. to analyze equivalence between the two multipoles representing the

translation of a single attack to attack.

To this purpose, referring again to Figure 2.28, we can identify the multipoles $\mathscr{M}_c = (AF^c, R^c_{INP}, R^c_{OUTP})$ with $AF^c = (\{X_{A,B}, Y_{A,B}\}, (X_{A,B}, Y_{A,B}))$, $R^c_{INP} = \{(A, X_{A,B})\} \cup \{(C_1, Y_{A,B}), \ldots, (C_n, Y_{A,B})\}$, $R^c_{OUTP} = \{(Y_{A,B}, B)\}$, and $\mathscr{M}_{sa} = (AF^{sa}, R^{sa}_{INP}, R^{sa}_{OUTP})$ with $AF^{sa} = (\{\overline{AB}\}, \emptyset)$, $R^{sa}_{INP} = \{(D_1, \overline{AB}), \ldots, (D_m, \overline{AB})\} \cup \{(C_1, \overline{AB}), \ldots, (C_n, \overline{AB})\}$, and $R^{sa}_{OUTP} = \{(\overline{AB}, B)\}$.

It is immediate to observe that the replacement of $\mathscr{M}_c$ with $\mathscr{M}_{sa}$ (or vice versa) is in general not legitimate: for instance, $\mathscr{M}_c$ and $\mathscr{M}_{sa}$ are in general not equivalent if one considers a labelling *Lab* such that $Lab(D_1) = \texttt{in}$ and $Lab(A) = \texttt{in}$. However, this labelling is clearly illegal in a context where D_1 attacks A. More generally, the labels of arguments $D_1, \ldots, D_m$ completely determine the label of A, thus one may check whether the replacement is contextually legitimate. So, for any semantics **S**, we are interested in showing that $\mathscr{M}_c$ and $\mathscr{M}_{sa}$ are **S**-equivalent with respect to $\mathfrak{L}^{\mathbf{S}}_{\mathscr{R}} \triangleq \{F_{\mathbf{S}}(AF{\downarrow_{Args}}, \mathscr{M}_c{}^{\text{outp}}, \mathbf{L}_c, R^c_{OUTP}) \mid \mathbf{L}_c \in \mathfrak{L}_{\mathscr{M}_c\text{outp}}\} \cup \{F_{\mathbf{S}}(AF{\downarrow_{Args}}, \mathscr{M}_{sa}{}^{\text{outp}}, \mathbf{L}_{sa}, R^{sa}_{OUTP}) \mid \mathbf{L}_{sa} \in \mathfrak{L}_{\mathscr{M}_{sa}\text{outp}}\}$.

Observe that since the output relation of both $\mathscr{M}_c$ and $\mathscr{M}_{sa}$ consists of a single attack (arising from $Y_{A,B}$ and $\overline{AB}$ respectively) the two multipoles are equivalent if and only if the labels assigned to $Y_{A,B}$ and $\overline{AB}$ are the same for any labelling in $\mathfrak{L}^{\mathbf{S}}_{\mathscr{R}}$. Moreover, we focus on **CO**-equivalence of multipoles without loss of generality: in fact, it turns out (and it is easy to see) that, in any case, for both AF^c and AF^{sa}, the local function of complete semantics prescribes exactly one complete labelling, which implies that this is the only labelling also for all the other complete-compatible semantics considered in this chapter.

A remark is now in order concerning the labellings of the arguments outside the multipoles, i.e. $\mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}}$ (note that $\mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}} \supseteq \mathfrak{L}^{\mathbf{S}}_{\mathscr{R}}$ for any complete-compatible semantics **S** considered in this chapter). As we are interested in proving an equivalence result whatever the remaining part of the framework is (in addition to the arguments depicted in Figure 2.28), the set of labellings $\mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}}$ can not be precisely characterized as it also depends on the (unspecified) remaining part of the framework. As a consequence, we prove a sligthly stronger result, considering any labelling in the set $\overline{\mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}}}$ consisting of the labellings compatible with the attacks from the arguments D_i to the argument A. Since $\overline{\mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}}} \supseteq \mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}}$ this implies the desired equivalence result.

Now, the examination of labellings in $\overline{\mathfrak{L}^{\mathbf{CO}}_{\mathscr{R}}}$ can be carried out considering nine cases, i.e., all possible combination, for the sets $\{D_1, \ldots, D_m\}$ and $\{C_1, \ldots, C_n\}$, of three basic cases: i) there is an argument labelled `in`; ii) all arguments are labelled `out`; iii) otherwise (note in particular that $Lab(A)$ is determined by the labelling of the set $\{D_1, \ldots, D_m\}$ in both multipoles). As all cases are rather simple, for the sake of compactness we synthesize the analysis in Table 2.3 rather than providing trivial and verbose explanations: by inspection of the last two columns it appears that $Lab(Y_{A,B}) = Lab(\overline{AB})$ in all cases, as desired.

We have thus proved that the replacement of the considered multipoles is contextually **S**-legitimate for any complete-compatible semantics **S** considered in this chapter. Then, the replacement is safe for any such semantics **S** which is strongly transparent with respect to these multipoles. Given that the multipoles are acyclic, from the results recalled in Table 2.2 it follows that the replacement is guaranteed to be safe for all semantics considered in this chapter, but **SST**, for which the answer is negative in general and the question is open for this specific case.

The lesson learned is twofold: first, we have given a substantial formal confirmation of the intuition that the two translation procedures are equivalent as far as the "external effects" are concerned for a comprehensive set of semantics, second we have seen however that even "simple" and basically correct intuitions require a careful semantics-specific scrutiny which may point out specific exceptions or critical issues (like for semi-stable semantics in our case).

$\{D_1,\ldots,D_m\}$	$\{C_1,\ldots,C_n\}$	$Lab(A)$	$Lab(X_{A,B})$	$Lab(Y_{A,B})$	$Lab\overline{AB}$
$\exists$in	$\exists$in	out	in	out	out
$\exists$in	$\forall$out	out	in	out	out
$\exists$in	$\nexists$in $\wedge$ $\exists$undec	out	in	out	out
$\forall$out	$\exists$in	in	out	out	out
$\forall$out	$\forall$out	in	out	in	in
$\forall$out	$\nexists$in $\wedge$ $\exists$undec	in	out	undec	undec
$\nexists$in $\wedge$ $\exists$undec	$\exists$in	undec	undec	out	out
$\nexists$in $\wedge$ $\exists$undec	$\forall$out	undec	undec	undec	undec
$\nexists$in $\wedge$ $\exists$undec	$\nexists$in $\wedge$ $\exists$undec	undec	undec	undec	undec

Table 2.3: Contextually **CO**-legitimate replacement of $\mathscr{M}_c$ with $\mathscr{M}_{sa}$.

2.11 Related work

As mentioned in Section 2.1, the work presented in this chapter has connections with three main (and non-disjoint) topics in the area of computational argumentation namely:

- local evaluation in argumentation semantics;
- argumentation dynamics;
- equivalence and interchangeability between argumentation frameworks.

We discuss the relationships with the relevant literature orderly in the following subsections.

Local evaluation in argumentation semantics

As to our knowledge the first analyses of semantics' properties exploitable for the purpose of local evaluation in the literature are provided by the work on SCC-recursiveness [245] and the notion of directionality introduced in [15].

Starting from the latter, in a nutshell a semantics is directional when it is guaranteed that, as far as extensions are concerned, a part of the framework which does not receive attacks from the rest of the framework is unaffected by the rest of the framework itself. Letting U be a set of arguments not receiving attacks from arguments not in U, this means that the same results (i.e. the same set of local extensions) are obtained either by computing the global extensions and then intersecting them with U, or by directly computing the extensions of the restricted framework consisting of the arguments in U and of the attacks among them.

Directionality allows for local computation when the results one is interested in can be obtained by focusing on an unattacked set, but has no embedded notion of progressive construction: it simply prescribes a relation of inclusion between the local extensions and the global ones. As such it is poorly related with the properties of semantics decomposability and transparency. To give some examples, stable semantics, which is not directional, is fully decomposable (and hence strongly transparent) while semi-stable semantics (which is non directional too) lacks any form of decomposability and transparency. Admissible and complete semantics are directional, fully decomposable and strongly transparent, while ideal semantics (which is directional too) lacks any form of decomposability and satisfies only a very weak form of transparency. To

complete the picture, recall that grounded and preferred semantics (which feature intermediate properties) are directional too.

The notion of SCC-recursiveness has closer relationships with the present work, as already evidenced by the fact that we considered partition selectors based on the notion of SCC. Basically, the SCC-recursive scheme provides a general method to build the global extensions prescribed by a semantics by proceeding progressively following the (partial) order among SCCs induced by the attack relation (recall that the graph obtained by considering each SCC as a single node is acyclic). The SCC-recursive scheme applies to each SCC a semantics-specific *base function* and then prescribes how to "propagate the effects" of the choices made in the previous SCCs to the subsequent ones before applying in turn the base function to them. As such, SCC-recursiveness directly implies the property of semantics decomposability with respect to the selector $\mathscr{F}_{\mathsf{SCC}}$.

Five of the semantics we have considered in this chapter are SCC-recursive (namely admissible, complete, stable, grounded, and preferred semantics), and indeed we have proved that all of them feature stronger decomposability properties than the one implied by SCC-recursiveness. Moreover, the notion of local function introduced in this chapter can be seen as a generalization of the notion of base function in the SCC-recursive scheme.

Drawing a more detailed analysis of the relationships of SCC-recursiveness with decomposability and transparency properties is an interesting line of future work. As a first note in this direction, we can observe that the two semantics lacking SCC-recursiveness considered in this chapter (namely semi-stable and ideal semantics) lack also any decomposability property.

In [272] the problem of combining local evaluations is addressed in a multi-agent scenario context where each agent owns a part of the framework and may locally adopt a different semantics. This gives rise to the notion of multi-sorted argumentation framework where a global argumentation framework is regarded as composed of a set of interacting *cells*, each associated with a (possibly) different semantics. In this context, the investigation in [272] follows a sort of top-down approach: given a (global) set of arguments S, it addresses the problem of checking whether S is an extension of the multi-sorted framework, according to local evaluations carried out for each cell. Basically, the definition of local evaluation at the cell level, directly reuses notions taken from the SCC-recursive scheme, as explicitly stated in [272]: the acceptance functions used at the cell level (Definition 5 in [272]) correspond to the base functions of the SCC-recursive scheme, while the notions of subframework and qualified arguments of a subframework (Definitions 7 and 8 in [272]) also have a direct correspondence with key technical elements of the SCC-recursive scheme (respectively with $AF{\downarrow}_{UP_{AF}(S,E)}$ and $U_{AF}(S,E)$ in Definition 20 of [245]). Thus, in a sense, the work of [272] reuses some of the main notions of the SCC-recursive scheme by applying them into two important directions of generalization: considering arbitrary (rather than SCC-based) partitions of the framework and allowing heterogeneous local evaluations. However, the direct reuse of notions specifically conceived in the context of the SCC-recursive scheme limits the possibility to fully encompass situations of mutual interaction and cyclic dependence between cells, which are impossible in the case of SCCs but are possible with arbitrary partitions. The present work addresses the study of homogeneous local evaluations for arbitrary partitions of an argumentation framework by introducing novel notions to capture the more complex interactions between subframeworks arising in this context. Extending the results presented in this chapter to the case of heterogeneous local evaluations is an important direction of future work.

It has also to be mentioned that some results concerning the use of the same semantics (or of semantics with common properties) in all cells are provided in [272] (in particular the notion of *Uniform Case Extension Equivalence* in Definition 10 of [272] roughly corresponds to our notion of semantics decomposability). These results are not directly comparable with ours, due to the different modeling of the interactions between

subframeworks mentioned previously. For instance, in Example 5 of [272] a counterexample is given disproving (a sort of) top-down decomposability of grounded semantics in multi-sorted frameworks, while in our context grounded semantics is actually top-down decomposable.

In the notion of conditional acceptance function introduced in [57], basically the acceptance function, corresponding to a given semantics, accepts as input not only an argumentation framework but also an (externally imposed) condition, which corresponds to the set of possible labellings of the framework. In other words, the acceptance function is constrained to produce a set of labellings which is a subset of the given condition. This expresses some form of external influence on argument evaluation, and in this sense could be related to our notion of argumentation framework with input. However, it is based on a rather different intuition, since it expresses a constraint on the labels of all arguments, independently of any attack relation coming from outside, while in our approach external influences manifest themselves through attack relations involving a well-identified set of arguments in the conditioned framework. In [57] the generic notion of conditional acceptance function is instantiated only for complete semantics, while its application to other semantics is, as to our knowledge, still to be developed.

Abstract dialectical frameworks (ADFs) [62] generalize Dung's framework by detaching the meaning of attack from the binary relation between arguments, so that each element of this relation is just a link representing a dependency. The meaning of the dependencies for each argument s is then expressed by an acceptance condition C_s which associates each subset of the set of parents of s with either in or out, namely gives a binary decision on the acceptance of s given the set of its parents which are accepted. Hence, in ADFs argument evaluation is, by definition, based on a strictly local criterion and any global evaluation arises bottom-up from the combination of the local ones.

While the present work is strictly focused on Dung's framework and the relevant semantics based on the attack relation, it appears that the basic ideas underlying our analysis have significant commonalities with the process of bottom-up evaluation in ADFs. Generalizing the results we have obtained to the context of ADFs is therefore a very important direction of future work.

Argumentation dynamics

Broadly speaking, in the context of abstract argumentation, dynamics concerns the evolution of a given framework to which one or more modifications (i.e. additions and/or deletions of arguments and/or attacks) are applied. These modifications can be exogenous and neutral, namely determined by some external event, or endogenous and goal-oriented, namely deliberately induced by an agent to reach some goal, like the acceptance of a desired argument. In the former case, the main interest is in determining the effect of the external modifications, in the latter, in identifying the minimal set of modifications sufficient to reach the goal. In both cases, one is typically interested in reusing as far as possible the results of previous computations carried out in the original framework so as to limit the amount of new computation required by the modification. Hence some of the pre-existing computation results have to be combined with the results of some partial computations in the new framework. Clearly the results presented in this chapter are specifically related to this facet of argumentation dynamics and we focus on the relevant literature. A detailed analysis of the broader implications of our work on argumentation dynamics is beyond the scope of the present chapter.

In [201] to save computation in a dynamic context the *division-based* method is proposed. Essentially, after a modification, the considered framework is divided into two parts, one unaffected and one affected. Briefly, the affected part consists of those arguments which are reachable (through a directed path of attacks) starting from any argument or attack involved in the modification. The identification of the unaffected part

relies on the directionality property, which is required for the application of the method. To formalize the influence of the unaffected part over the affected part, the notion of conditioned argumentation framework is introduced, namely an argumentation framework receiving some attacks from arguments included in another argumentation framework. The chapter then deals with incremental computation for some semantics satisfying the directionality property, namely complete, grounded, preferred and ideal[14] semantics. After the modification, one needs to recompute the extensions only for the affected part (modeled as a conditioned argumentation framework w.r.t. the unaffected part).

Some basic notions underlying the division-based method are related to our work. In particular, the notion of conditioned argumentation framework in [201] is similar to the notions of conditioning relation and of argumentation framework with input in our Definitions 12 and 13. Moreover, the incremental computation in a conditioned framework is analogous to the application of the local function introduced in Definition 13.

There is however an important difference due to the fact that the division-based method is essentially based on the directionality principle and, in particular, requires that there are no paths from the affected part to the unaffected part. As a consequence, the division-based method covers the cases where a framework is partitioned into two subframeworks such that one has an output, without having an input, and the other has an input (from the former) without having an output. The results concerning incremental computation of the four semantics considered in [201] correspond to a restricted form of semantics decomposability under these restrictive assumptions: both the unaffected and the affected part consist of a set of SCCs such that the SCCs included in the unaffected part precede those included in the affected part according to the partial order induced by the attack relation. Given this observation, the notion of decomposability in this context basically corresponds to a mild generalization of decomposability w.r.t.. $\mathscr{F}_{\mathsf{SCC}}(AF)$, i.e. of the weakest notion of decomposability considered in this chapter, and is definitely weaker than decomposability w.r.t.. $\mathscr{F}_{\cup\mathsf{SCC}}(AF)$. Our work is definitely more general as it concerns arbitrary partitions and does not rely on the directionality property. In particular, we prove full decomposability of stable semantics, which is not directional.

The work on splitting argumentation frameworks [29] focuses on modifications involving only additions of arguments or attacks (called expansions) and, apart of this restriction, shares the main basic assumptions with [201]. Considering a subclass of expansions called weak expansions, a splitting divides an argumentation framework into two subframeworks, such that only one of them receives attacks from the other: the two subframeworks correspond to the unaffected and affected parts of [201]. To model the effect of the unaffected subframework on the affected one, in [29] a modification of the affected subframework is introduced, which involves the addition of self-attacks and bears some similarity with our notion of standard argumentation framework for an argumentation framework with input. Then, the *splitting theorem* of [29] provides a decomposability result for stable, admissible, preferred, complete and grounded semantics, which, due to the restriction on the partitions considered, as in the case of [201], are weaker than the ones considered in this chapter.

The restrictions that one of the two parts can not receive attacks from the other one is lifted in [30] where an arbitrary partition of a framework into two parts is called *quasi splitting* and, using a technical arsenal rather different than ours, the decomposability property of stable semantics is proved. We achieved the same result for stable semantics in the context of a more general analysis, covering six additional literature semantics.

On the performance side, there are some empirical evidences that both the division-based method [200] and the splitting approach [28] may significantly reduce the computation time required for some standard

[14] Actually the claim concerning ideal semantics turns out to be flawed, as recently pointed out in [18].

problems in abstract argumentation w.r.t. to algorithms adopting a "monolithic" approach. Investigating the advantages provided by our more general approach in this respect is an important direction of future work.

Equivalence and interchangeability between argumentation frameworks

Various notions of equivalence for argumentation frameworks have been considered in the literature. The most basic ones focus either on structural correspondences (like the notion of syntactical equivalence, i.e. equality of arguments and attacks, used in [240] or the notion of isomorphism used in [15]) or on equality of extensions (w.r.t. a given semantics), which is called equivalence *tout court* in [240] and is analogous to the notion of equivalence between logic programs [203]. These notions are poorly or not at all related with modularity and interchangeability issues, that may arise in various contexts and in particular in presence of some form of argumentation dynamics.

To address this limitation, the notion of strong equivalence between argumentation framework (again, analogous to the one of strong equivalence between logic programs [203]) is introduced and investigated in [240]: two frameworks F and G are strongly equivalent w.r.t. a given semantics if for any argumentation framework H, the frameworks $F \cup H$ and $G \cup H$ have the same extensions. Basically, F and G must preserve the same outcomes in front of any operation of expansion. Since this requirement is, in fact, very strong, weaker notions of equivalence have subsequently been considered in the literature by restricting the set of expansions of the original frameworks encompassed. In particular four subclasses of expansions (called normal, weak, strong, and local[15]) are considered in [25] giving rise to four correspondent definitions of expansion equivalence all weaker than strong equivalence. A different notion of equivalence, introduced in [26], refers to the problem of minimal change: given a framework and a set of arguments E whose (credulous) acceptance has to be enforced, one is interested in identifying the minimal number of modifications that ensure the desired enforcement result. Two frameworks are minimal change equivalent, if for any set E the minimal number of modifications required to enforce E is the same in both frameworks. The relationships between all the above mentioned notions of equivalence have been analyzed in detail in [27].

The approach presented in this chapter is complementary to the ones reviewed above: while these refer to several forms of invariance over the whole framework w.r.t. an operation of expansion, our work concerns invariance only in the unmodified part of the framework w.r.t. an operation of replacement. This involves a notion of equivalence in terms of Input/Output behavior and the study of the property of semantics transparency, which have no counterpart in the works cited above. As already mentioned, they can be related with the notion of strong equivalence in logic programming, while our approach is closer in spirit to the notion of modular equivalence between logic programs [239] and, more generally, with the study of modularity in this context [177]. A detailed analysis of the possible interplay between our results and the area of modular logic programming is beyond the scope of the present chapter and is left for future work.

The issue of substitutions within an argumentation framework is explicitly addressed by the notion of *fibring* [149] which indeed covers the more general case of combining together networks of different nature (e.g. embedding a neural network or a Bayesian network into an argumentation framework), including the special case of combination of networks of the same nature, called *self-fibring*. Due to the potential heterogeneity of the networks involved, however, fibring concerns the substitution of a *single node* of a network with an entire other network (neither of them having a notion of "interface" with the rest of the framework) and hence addresses a different kind of replacement than the one considered in this chapter, which involves

[15] This terminology, taken out of its context, may be a bit misleading: normal expansions are not the most general case of expansions, and the terms weak and strong here refer to the additional arguments, so strong expansions are not a subset of weak expansions.

the two argumentation multipoles, i.e. two partial networks with well-defined interface. Moreover, the study presented in [149] covers generalized argumentation frameworks, featuring a richer set of relations (e.g. support, attacks to attacks, attacks arising from attacks, collective and disjoint attacks) than Dung's framework, and investigates how this conceptual and technical arsenal can be used to properly transform the incoming and outcoming links involving just one node into links involving the nodes of the network replacing that node. Thus, the analysis in [149] goes deeply into these complex structural manipulations, which are mostly semantics independent, and does not concern the study of specific semantics properties. Our work, as already mentioned, concerns a different kind of replacement and lies in the context of traditional Dung frameworks, where we provide a systematic assessment of interchangeability-related properties for a comprehensive set of literature semantics. Extending and relating our results to generalized frameworks in the spirit of [149] is a further interesting direction of future work.

In [308] the notion of argumentation pattern is introduced in order to capture "general reusable solutions to commonly occurring problems in the design of argumentation frameworks". Hence an argumentation pattern is understood as a reusable and modular component, in a spirit which has some analogy with the idea of argumentation multipole introduced in this chapter. It has however to be observed that the notion of argumentation pattern lies at a higher level of abstraction than the one of argumentation multipole: the definition of argumentation pattern given in [308] involves a set of arguments and, basically, a set of possible labellings of these arguments. No notion of attack is explicitly involved, since an argumentation pattern captures a set of evaluation outcomes which together represent a "typical situation" seen from outside, independently of the (in fact, not necessarily univocal) underlying structure giving rise to this situation. Indeed, in [308] methods to translate (or flatten) a pattern into an argumentation framework and vice versa to extract a pattern from an argumentation framework (where arguments to be included in the pattern have been preliminarily identified) are devised. Our work, lying at different level, provides suitable technical foundations for further developments of the study of argumentation patterns. Indeed, our analysis concerning the equivalence of alternative representations of attacks to attacks in Section 2.10 strenghtens the analysis of patterns for so called higher-order attacks in Section 3.2 of [308]. Moreover in [308] the issue of pattern combination is mentioned as a matter of future research, which may certainly benefit from the systematic set of results provided in this chapter, applicable to the underlying flattened representation.

2.12 Conclusion

This chapter contributes to the emerging research direction on modularity-based properties and techniques in abstract argumentation, by introducing a novel comprehensive formal corpus to describe the Input/Output behavior of argumentation frameworks along with the relevant semantics properties, and by providing a systematic assessment of seven well-known argumentation semantics in this context. Due to their foundational nature, we believe these results may play an enabling role in the development of a variety of more specific investigation lines, ranging from the sound combination of heterogeneous semantics to the definition of reusable argumentation patterns. As to future work, in addition to the many issues already included in the discussion of Section 2.11, we mention three further interesting lines. First, the extension of the analysis carried out in the chapter to other literature semantics like the ones mainly based on the notion of conflict-freeness (e.g. stage [306], CF2 [17, 245], stage2 [152] semantics) or those featuring a parametric definition (e.g. resolution-based semantics [14]). Second, the study of *argumentation synthesis* problems, namely, given a desired Input/Output behavior generating an argumentation framework which produces it, possibly under some constraints concerning its structure and/or the semantics to be adopted. Third, a systematic

definition of modularity-related variations of traditional computational problems in abstract argumentation, e.g. checking whether two multipoles are equivalent according to a given semantics, and the analysis of their complexity properties.

Chapter 3

Reasoning about trust through argumentation

3.1 Introduction

This chapter synthesizes my contributions in the area of computational models of argument, dealing, more precisely, with the use of abstract argumentation to reason about trust in multiagent systems. These contributions have been published in several venues:

- *Serena Villata, Guido Boella, Dov M. Gabbay, Leendert van der Torre. Arguing about the Trustworthiness of the Information Sources. 11th European Conference Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2011): 74-85* [310],
- *Célia da Costa Pereira, Andrea Tettamanzi, Serena Villata. Changing One's Mind: Erase or Rewind? 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011): 164-171* [112],
- *Célia da Costa Pereira, Mauro Dragoni, Andrea Tettamanzi, Serena Villata. Fuzzy Labeling for Abstract Argumentation: An Empirical Evaluation. 10th International Conference on Scalable Uncertainty Management (SUM 2016): 126-139* [114],
- *Fabio Paglieri, Cristiano Castelfranchi, Célia da Costa Pereira, Rino Falcone, Andrea Tettamanzi, Serena Villata. Trusting the messenger because of the message: feedback dynamics from information quality to source evaluation. Computational & Mathematical Organization Theory 20(2): 176-194 (2014)* [247],
- *Serena Villata, Guido Boella, Dov M. Gabbay, Leendert van der Torre. A socio-cognitive model of trust using argumentation theory. Int. J. Approx. Reasoning 54(4): 541-559 (2013)* [309].

The contributions reported in this chapter are the result of my fruitful collaboration with Celia da Costa Pereira (UNS) and Andrea Tettamanzi (UNS). They also resulted from a collaboration with Fabio Paglieri (CNR Rome), Rino Falcone (CNRS Rome) and Cristiano Castelfranchi (CNR Rome). Part of the contributions presented in this chapter originated from the results I published in my PhD thesis, i.e., the meta-argumentation methodology here applied to reasoning about conflicts in trust. Finally, this line of work has been continued in the PhD thesis of Amel Ben Othmane that I supervised with Andrea Tettamanzi and Nhan

Le Than. She successfully defended her PhD thesis titled "CARS - A multi-agent framework to support the decision making in uncertain spatio-temporal real-world applications" on October 2017.

Trust is a mechanism for managing uncertain information in decision making, taking into account also the sources besides the content of information only. In their interactions, agents have to reason to decide whether they should trust or not the other sources of information, and on the extent to which they trust those other sources. This is important, for example, in medical contexts, where doctors have to inform the patient of the pro and con evidence from different sources concerning some treatment, in decision support systems where the user is not satisfied by an answer without explanations, or in trials where judges have to specify the motivations about which conflicting evidence they trust. A cognitive analysis of trust is fundamental to predict very different strategies for building or increasing trust, for founding mechanisms of reputation, persuasion, and argumentation in trust building [90]. The contribution of this chapter is twofold: on the one side, we present an argumentation-based approach to reason on trust conflicts using the methodology of *meta-argumentation* I introduced in my PhD thesis [307], and, on the other hand, we present a fuzzy labeling algorithm to label arguments in an argumentation framework based on the trustworthiness of the source proposing them.

The chapter is organized as follows: Section 3.2 describes the challenges of reasoning about trust conflicts using meta-argumentation and presents the proposed formal framework, and Section 3.3 describes the fuzzy labeling algorithm and how it can be employed in a belief revision scenario in a multiagent system. Section 3.4 compares the proposed approaches to reason about trust using argumentation theory to the related literature, and conclusions end the chapter.

3.2 A cognitive model of conflicts in trust using argumentation

In this section, we start from the cognitive model of trust introduced by Castelfranchi and Falcone [90], and we present a cognitive model of conflicts in trust using argumentation. In particular, the reasoning process addressed by the agents concerning the extent to which they trust the other information sources leads to the emergence not only of conflicts among the information but also of the conflicts among the sources. Since argumentation is a mechanism to reason about conflicting information [127] it seems the suitable methodology for reason about trust. When two pieces of information coming from different sources are conflicting, they can be seen as two arguments attacking each other. When an information source explicitly expresses a negative evaluation of the trustworthiness of another source, it can be seen as an "attack" to the trustworthiness of the second source modelled as an argument as well. To deal with the dimension of conflict in handling trust, we propose to use argumentation theory, modelling both information and information sources as arguments and arguing about them. In argumentation theory [266], the arguments are considered to be accepted or not depending on the attacks against them. In standard argumentation frameworks, neither the information sources proposing the arguments nor their trustworthiness are considered. In recent years, the area has seen a number of proposals [261, 293, 216, 253, 310, 113] to introduce the trust component in the evaluation process of the arguments. The common drawback of these approaches is that they do not return the intrinsic complexity of the trust notion, as highlighted instead by socio-cognitive models like [90].

The challenge of this work is to use argumentation theory not only to model whether an information source is trusted or not, but also to understand the reasons, modeled under the form of arguments, for trusting the sources in case of conflicts concerning their trustability. This means that we need to distinguish the conflicts about the content of the arguments which are usually specified through an attack relation, and the conflicts about the different opinions of the sources on the trustworthiness of the other sources. These are

two separate reasoning levels, and the challenge is to model both of them using argumentation theory. In particular, we present a way to deal with the conflicts about trust using Dung's abstract argumentation framework [133]. It is not obvious how to model in a Dung argumentation framework the trust about arguments and the conflicts about sources' trustworthiness. A Dung argumentation framework can be instantiated by the arguments and attacks defined by a knowledge base. The knowledge base inferences are defined in terms of the claims of the justified arguments, e.g., the ASPIC+ framework [263] instantiates Dung frameworks with accounts of the structure of arguments, the nature of attack and the use of preferences. In such a kind of framework, arguments are instantiated by sentences of a single knowledge base, without reference to the information sources. The only possibility is to include sources and trust inside the content of the argument. This makes it difficult to distinguish between the object level concerning content of information and the meta-level concerning trust, sources and the conflicts among them. In reasoning about trust, the information about the trustworthiness relations among the sources are meta-level information, and they cannot be inserted directly into the argumentation framework. They influence the behavior of the framework in the sense that they lead to further conflicts among the sources and their information items, i.e., what the sources claim.

The following example presents informally the opinions of several witnesses during a trial, illustrating conflicts about trust among the sources and not only among the pieces of information they provide, where the external evaluator is the judge:

- *Witness1: I suspect that the man killed his boss in Rome. (a)*
- *Witness1: But his car was broken, thus he could not reach the crime scene. (b)*
- *Witness2: Witness1 is a compulsive liar. (c)*
- *Witness3: I repaired the suspect's car at 12pm of the crime day. (d)*
- *Witness4: I believe that Witness3 is not able to repair that kind of car. (e)*
- *Witness5: The suspect has another car. (f)*
- *Witness6: Witness5 saw that the suspect parked 2 cars in my underground parking garage 3 weeks ago. (g)*
- *Witness2: Witness5 was on holidays 3 weeks ago. (h)*
- *Witness7: Witness5 cannot go on holidays because of his working contract. (i)*
- *Witness3: Witness7 is not competent about the working contracts of the underground parking garage. (l)*
- *Witness1: Witness7 does not really think that Witness5 cannot go on holidays because of his working contract. (m)*

In these sentences, different kinds of conflicts are highlighted among the sources concerning their trustability. What we call the object level is illustrated by the arguments (a) and (b): Witness1 would believe the suspect is the murderer but he explains that another argument (the car was broken) prevents this conclusion. Thus argument (b) attacks argument (a) since they are conflicting. But attacks can concern also the trustability of sources, once this aspect is modelled in terms of arguments (meta-arguments) as well. First, the sources can attack the trustworthiness of the other sources, see, e.g., argument (c) attacking the trustworthiness of Wintess1. Second, we must model the connection between the argument about the trustability of Witness1 and the arguments (a) and (b) - as well as the attack between the two arguments - he advances. The sources must be modelled as evidence motivating their arguments or attacks, which otherwise should be considered as unacceptable. Moreover, sources can provide evidence also concerning the other sources' arguments, e.g., argument (g) provides evidence for argument (f). Third, while attacks like the one done

by argument (c) are addressed against the sources' trustworthiness as a whole (represented through a meta-argument), conflicts about trust can be restricted to a particular argument or attack proposed by a source who is not considered untrustworthy in general. E.g., argument (h) expresses concerns about the trustworthiness of argument (g) and not about the source itself. Fourth, conflicts about the trustworthiness of the sources can be further detailed in order to deal with the competence of the sources, e.g., argument (l), and their sincerity, e.g., argument (m). The example leaves as implicit the issue of a feedback between the trustworthiness of the information items and the sources' trustworthiness when what they said is attacked.

As mentioned before, in standard argumentation frameworks [266] it is difficult to formalize the example above with sentences from a single knowledge base only, e.g., to model it in ASPIC+ style instantiated argumentation. Moreover, meta-level information such as the distinction about conflicts based on sincerity and those based on competence cannot be represented in those frameworks. These two trust dimensions might be independently evaluated in the argumentation process: Bob's sincerity/honesty (Alice believes that Bob has told her the truth) vs. Bob's competence (Alice trusts the judgment of Bob if he is expert). Finally, it has to be modeled the fact that attacking Bob's argument means attacking Bob and his credibility and trustworthiness as source. This is fundamental, both in the case in which it is intentional and it is the real objective of the attack, or when it is not intended but is a consequence of the invalidation of the arguments. This is because of the bidirectional link between the source and its information items: the provided item is more or less believable on the basis of the source trustworthiness, but the invalidation of the item feedbacks on the source's credibility.

In this section, we address the following research question:

- How to model the socio-cognitive aspects of trust using argumentation theory?

The research question breaks down into the following subquestions:

1. How to represent the information sources and attack their trustworthiness?
2. How to represent pro and con evidence, as done in Carneades [156]?
3. How to attack the sources' trustworthiness about single information items?
4. How to represent the trust feedback between the sources and their information items?
5. How to distinguish the two dimensions of trust, i.e., sincerity and competence?

To answer the research questions, we propose meta-argumentation [176, 230, 54, 93, 307]. Meta-argumentation provides a way to instantiate abstract arguments, i.e., abstract arguments are treated as meta-arguments: arguments about other arguments. It allows us not only to reason about arguments such as sentences from a knowledge base indexed by the information source, but also to introduce in the framework, at the meta-level, other instances like arguments about the trustworthiness of sources. The advantage of adopting meta-argumentation is that we do not extend Dung's framework in order to introduce trust but we instantiate his theory with meta-arguments. For a further discussion about meta-argumentation, see Villata [307].

The sources are introduced into the argumentation framework under the form of meta-arguments of the kind "*agent i is trustable*". An attack to the trustworthiness of a source is modeled as an attack to the meta-argument "*agent i is trustable*". Similarly, in meta-argumentation, both arguments and attacks are represented as meta-arguments, thus allowing arguments to attack attacks.

Each source motivates the information items it proposes via meta-arguments which represent the need of evidence to make an argument acceptable. Each argument simply "put on the table" is considered unacceptable if no sources provide an evidence motivating it by being considered trustable. We show that the property such that an argument which is not motivated by evidence is not accepted holds in our model.

The information sources propose information items, i.e., arguments, and attacks among these arguments. An attack to the trustworthiness of an item or an attack is modeled as an attack in the meta-level to the evidence provided by the source for that item. We prove that in our model it holds that if there is only one untrustworthy source showing evidence in favor of an argument then this argument cannot be accepted.

The feedback from the sources to the information items and back is modeled again by introducing new meta-arguments, and the attacks among them. These meta-arguments model a sort of threshold such that if a number of attackers of the information items proposed by a source are accepted, i.e., trustable, thus the attacked source cannot be considered trustworthy.

Finally, the two dimensions of sincerity and competence are modeled using a meta-argument of the kind "*a is believed by source i*" representing the fact that argument a is believed by the source, and thus the source is sincere in proposing a. This meta-argument supports the "real" meta-argument which models argument a in the meta-level. An attack towards the source's sincerity is modeled as an attack towards the meta-argument representing the believed argument while an attack to the competence is directed towards the motivation the "believed" meta-argument provides to the "content" meta-argument. We show that in our model it holds that if a trustworthy source attacks the trustworthiness of argument, then the extensions are the same if it attacks the sincerity or the competence about the argument.

Note that we do not claim that argumentation is the only way to model trust, but we underline that, when the sources argue, they are strongly influenced by the trustworthiness they assign to the other sources. Moreover, we do not assign a numerical value associated to trust, because we are more interested in reasoning about the motivations of the sources, e.g., in the case of Witness1 we have that he explains that he does not believe a and that this is due to argument b. Finally, we do not treat converging and diverging beliefs sources, and the source's subjective uncertainty [90]. This is left as future work.

Meta-argumentation

Meta-argumentation instantiates Dung's theory with meta-arguments, such that *Dung's theory is used to reason about itself* [53, 54, 307]. Meta-argumentation is a particular way to define mappings from argumentation frameworks to extended argumentation frameworks: arguments are interpreted as meta-arguments, of which some are mapped to "argument a is accepted", $acc(a)$, where a is an abstract argument from the extended argumentation framework (EAF). Moreover, auxiliary arguments are introduced to represent, for example, attacks, so that, by being arguments themselves, they can be attacked or attack other arguments. The meta-argumentation methodology is summarized in Figure 1.

Like Baroni and Giacomin [15], we use a function $\mathscr{E}$ mapping an argumentation framework $\langle Ar, \rightarrow \rangle$ to its set of extensions, i.e., to a set of sets of arguments. Since they do not give a name to the function $\mathscr{E}$, and it maps argumentation frameworks to the set of accepted arguments, we call $\mathscr{E}$ the *acceptance function*.

Definition 42. Let $\mathscr{U}$ be the universe of arguments. An acceptance function $\mathscr{E} : 2^{\mathscr{U}} \times 2^{\mathscr{U} \times \mathscr{U}} \rightarrow 2^{2^{\mathscr{U}}}$ is a partial function which is defined for each argumentation framework $\langle Ar, \rightarrow \rangle$ with finite $Ar \subseteq \mathscr{U}$ and $\rightarrow \subseteq Ar \times Ar$, and maps an argumentation framework $\langle Ar, \rightarrow \rangle$ to sets of subsets of Ar: $\mathscr{E}(\langle Ar, \rightarrow \rangle) \subseteq 2^{Ar}$.

The function f assigns to each argument a in the EAF, a meta-argument "argument a is accepted" in the basic argumentation framework. The function f^{-1} instantiates an AF with an EAF. We use Dung's

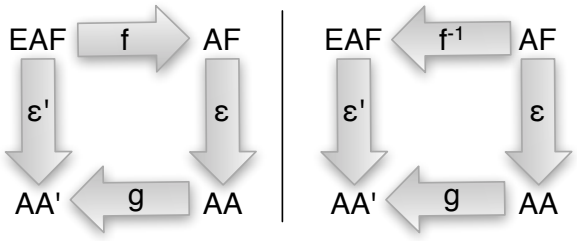


Figure 3.1: The meta-argumentation methodology workflow.

acceptance functions $\mathscr{E}$ to find functions $\mathscr{E}'$ between EAFs and the acceptable arguments AA' they return. The acceptable arguments of the meta-argumentation framework are a function of the extended argumentation framework: $AA' = \mathscr{E}'(EAF)$. The transformation function consists of two parts: the function f^{-1}, transforming an argumentation framework to an extended argumentation framework, and a function g which transforms the acceptable arguments of the argumentation framework into acceptable arguments of the extended argumentation framework. Summarizing, $\mathscr{E}' = \{(f^{-1}(a), g(b)) \mid (a,b) \in \mathscr{E}\}$ and $AA' = \mathscr{E}'(EAF) = g(AA) = g(\mathscr{E}(AF)) = g(\mathscr{E}(f(EAF)))$.

The first step of the meta-argumentation approach is to define the set of extended argumentation frameworks. The second step consists of defining flattening algorithms as a function from this set of EAFs to the set of all basic AF: $f : EAF \to AF$. The inverse of the flattening is the instantiation of the argumentation framework. See [54, 307] for further details. We define an EAF as a set of partial argumentation frameworks of the sources $\langle Ar, \langle Ar_1, \to_1 \rangle, \ldots, \langle Ar_n, \to_n \rangle, \to \rangle$ [116].

Definition 43. An extended argumentation framework (EAF) is a tuple $\langle Ar, \langle Ar_1, \to_1 \rangle, \ldots, \langle Ar_n, \to_n \rangle, \to \rangle$ where for each source $1 \leq i \leq n$, $Ar_i \subseteq Ar \subseteq \mathscr{U}$ is a set of arguments, $\to$ is a binary attack relation on $Ar \times Ar$, and $\to_i$ is a binary relation on $Ar_i \times Ar_i$. The universe of meta-arguments is $MU = \{acc(a) \mid a \in \mathscr{U}\} \cup \{X_{a,b}, Y_{a,b} \mid a,b \in \mathscr{U}\}$, where $X_{a,b}, Y_{a,b}$ are the meta-arguments corresponding to the attack $a \to b$. The flattening function f is given by $f(EAF) = \langle MA, \longmapsto \rangle$, where MA is the set of meta-arguments and $\longmapsto$ is the meta-attack relation. For a set of arguments $B \subseteq MU$, the unflattening function g is given by $g(B) = \{a \mid acc(a) \in B\}$, and for sets of subsets of arguments $AA \subseteq 2^{MU}$, it is given by $g(AA) = \{g(B) \mid B \in AA\}$.

Given an acceptance function $\mathscr{E}$ for an AF, the extensions of accepted arguments of an EAF are given by $\mathscr{E}'(EAF) = g(\mathscr{E}(f(EAF)))$. The derived acceptance function $\mathscr{E}'$ of the EAF is thus $\mathscr{E}' = \{(f^{-1}(a), g(b)) \mid (a,b) \in \mathscr{E}\}$. We say that the source i provides evidence in support of argument a when $a \in Ar_i$, and that the source i supports the attack $a \to b$ when $a \to b \in \to_i$.

Note that the union of all the Ar_i does not produce Ar because Ar contains also those arguments which are not supported by the sources, and are just "put on the table". Definition 44 presents the instantiation of a basic argumentation framework as a set of partial argumentation frameworks of the sources using meta-argumentation.

Definition 44. Given an $EAF = \langle Ar, \langle Ar_1, \to_1 \rangle, \ldots, \langle Ar_n, \to_n \rangle, \to \rangle$ where for each source $1 \leq i \leq n$, $Ar_i \subseteq Ar \subseteq \mathscr{U}$ is a set of arguments, $\to \subseteq Ar \times Ar$, and $\to_i \subseteq Ar_i \times Ar_i$ is a binary relation over Ar_i. $MA \subseteq MU$ is $\{acc(a) \mid a \in Ar_1 \cup \ldots \cup Ar_n\}$, and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that:

$acc(a) \longmapsto X_{a,b}, X_{a,b} \longmapsto Y_{a,b}, Y_{a,b} \longmapsto acc(b)$ if and only if there is a source $1 \leq i \leq n$ such that $a, b \in Ar_i$ and $a \rightarrow b \in \rightarrow_i$.

Intuitively, the $X_{a,b}$ auxiliary argument means that the attack $a \rightarrow b$ is "inactive", and the $Y_{a,b}$ auxiliary argument means that the attack is "active". An argument of an *EAF* is acceptable if and only if it is acceptable in the flattened argumentation framework.

Modelling trust in meta-argumentation

In this section, we formally define our cognitive model of trust using meta-argumentation. Using the running example described in the introduction, we show how the model can be used to formally model it, and we present some desired properties of our model.

Information sources. The reason why abstract argumentation is not suited to model trust is that an argument, if it is not attacked by another acceptable argument, is considered acceptable. This prevents us from modeling the situation where, for an argument to be acceptable, it must be related to some trusted sources which provide the evidence for such an argument to be accepted. Without an explicit representation of the sources, it becomes impossible to talk about trust: the argument can only be attacked by conflicting information, but it cannot be made unacceptable due to the lack of trust in the source.

Modelling evidence is another challenge: sources are a particular type of evidence. Arguments needing evidence are well known in legal argumentation, where the notion of burden of proof has been introduced [156]. Meta-argumentation provides a means to model burden of proof in abstract argumentation without extending argumentation. The idea is to associate to each argument $a \in Ar$ put on the table, which is represented by means of meta-argument $acc(a)$, an auxiliary argument $W_{acc(a)}$ attacking it. Being auxiliary this argument is filtered out during the unflattening process. This means that without further information, just as being "put on the table", argument a is not acceptable since it is attacked by the acceptable argument $W_{acc(a)}$, and there is no evidence defending it against this "default" attack, as visualized in Figure 3.2 for arguments a and b. In the figures, we represent the meta-arguments associated to the information sources as boxes, and the arguments as circles where grey elements are the acceptable ones. This evidence is modeled by means of the attacks towards these auxiliary arguments, e.g., $W_{acc(a)}$, leading to a reinstatement of meta-argument $acc(a)$. Attacks are modeled as arguments as well, so they need evidence to be acceptable. For each auxiliary argument $Y_{a,b}$, representing the activation of the attack, we associate an auxiliary argument $W_{Y_{a,b}}$.

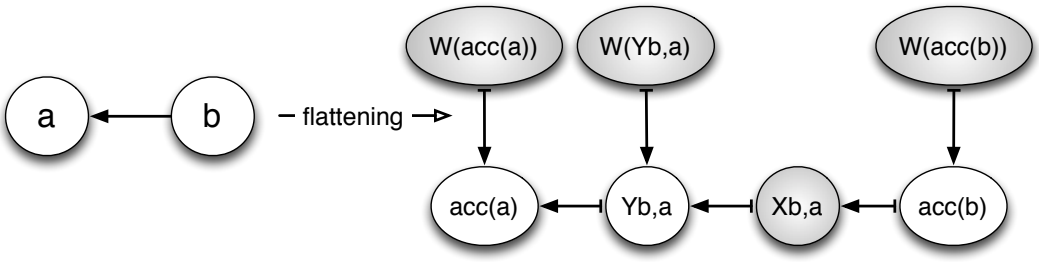


Figure 3.2: Arguments and attacks without evidence.

Sources are introduced in the meta-argumentation framework under the form of meta-arguments "*source s is trustable*", $trust(s)$, for all the sources s. Each argument a in the sources' mind is motivated by means of an attack on $W_{acc(a)}$. We represent the fact that one or more information sources provide evidence for the same argument by letting them attack the same $W_{acc(a)}$ auxiliary argument. An example of multiple evidence is depicted in Figure 3.3. As for arguments, an attack to become active needs some trusted agent.

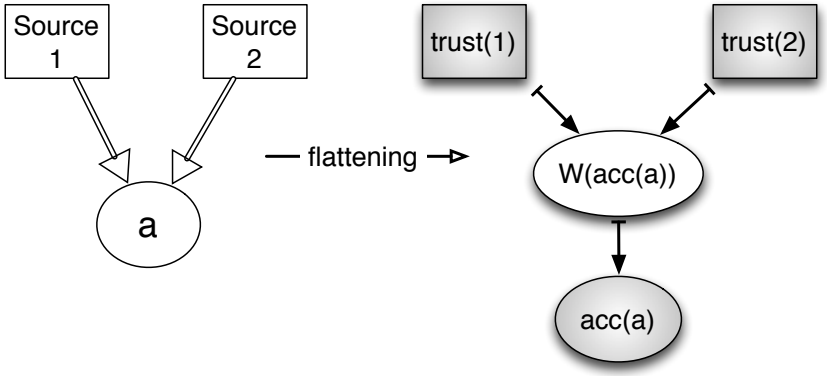


Figure 3.3: An example of multiple evidence.

Notice that the assumption that there must be evidence for an argument to be accepted is a very general and often used reasoning pattern, e.g., in causal reasoning, where everything needs to be explained, i.e., to have a cause / to be caused, as in the Yale shooting problem for instance. For more details about causal reasoning, see Bochman [52].

We have now to discuss which semantics we adopt for assessing the acceptability of the arguments and the sources. For example, suppose that two sources claim they are each untrustworthy. What is the extension? We adopt admissibility based semantics, i.e., $\Gamma \in \mathscr{E}_{\text{admiss}}(AF)$. We do not ask for completeness because if one wants to know whether a particular argument is acceptable, the whole model is not needed, just the part related to this particular argument is needed.

The reader should not be confused by the similarity between evidence and support [55]. The meaning of Boella et al. [55]'s notion of support is that if argument a is acceptable then argument b is acceptable too. Note that the supported argument b is acceptable (if not attacked) even without the support of a, i.e., a is not acceptable. Support exploits an auxiliary argument Z, but with some difference with the auxiliary argument W. First, given a supporting b, there is a $Z_{a,b}$ such that b attacks $Z_{a,b}$ and $Z_{a,b}$ attacks a, while, here, $W_{acc(a)}$ attacks the argument needing evidence. Second, there is a Z meta-argument for each supporting argument, while, here, there is only one W meta-argument attacked by all the arguments and agents giving an evidence. For more details about this model of support in argumentation, see Boella et al. [55].

We extend the definition of EAF (Definition 43) by adding evidence provided by the information sources and second-order attacks, such as attacks from an argument or attack to another attack. For more details about second-order attacks in meta-argumentation, see [230, 54]. The unflattening function g and the acceptance function $\mathscr{E}'$ are defined as above.

Definition 45. A trust-based extended argumentation framework $TEAF^2$ with second-order attacks is a tuple $\langle Ar, \langle Ar_1, \rightarrow_1, \rightarrow_1^2 \rangle, \ldots, \langle Ar_n, \rightarrow_n, \rightarrow_n^2 \rangle, \rightarrow \rangle$ where for each source $1 \leq i \leq n$, $Ar_i \subseteq Ar \subseteq \mathscr{U}$ is a set of arguments, $\rightarrow \subseteq Ar \times Ar$, $\rightarrow_i$ is a binary relation on $Ar_i \times Ar_i$, $\rightarrow_i^2$ is a binary relation on $(Ar_i \cup \rightarrow_i) \times \rightarrow_i$.

Definition 46 presents the instantiation of a $TEAF^2$ with second-order attacks as a set of partial frameworks of the sources using meta-argumentation.

Definition 46. Given a $TEAF^2 = \langle Ar, \langle Ar_1, \rightarrow_1, \rightarrow_1^2 \rangle \ldots, \langle Ar_n, \rightarrow_n, \rightarrow_n^2 \rangle, \rightarrow \rangle$, the set of meta-arguments MA is $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in Ar_1 \cup \ldots \cup Ar_n\} \cup \{X_{a,b}, Y_{a,b} \mid a,b \in Ar_1 \cup \ldots \cup Ar_n\} \cup \{W_{acc(a)} \mid a \in Ar_1 \cup \ldots \cup Ar_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that:

- $acc(a) \longmapsto X_{a,b}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $X_{a,b} \longmapsto Y_{a,b}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $Y_{a,b} \longmapsto acc(b)$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and
- $trust(i) \longmapsto W_{acc(a)}$ iff $a \in Ar_i$, and $W_{acc(a)} \longmapsto acc(a)$ iff $a \in Ar$, and
- $trust(i) \longmapsto W_{Y_{a,b}}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $W_{Y_{a,b}} \longmapsto Y_{a,b}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and
- $acc(a) \longmapsto X_{a,b \rightarrow c}$ iff $a,b,c \in Ar_i$ and $a \rightarrow_i^2 (b \rightarrow_i c)$, and $X_{a,b \rightarrow c} \longmapsto Y_{a,b \rightarrow c}$ iff $a,b,c \in Ar_i$ and $a \rightarrow_i^2 (b \rightarrow_i c)$, and $Y_{a,b \rightarrow c} \longmapsto Y_{b,c}$ iff $a,b,c \in Ar_i$ and $a \rightarrow_i^2 (b \rightarrow_i c)$, and
- $Y_{a,b} \longmapsto Y_{c,d}$ iff $a,b,c \in Ar_i$ and $(a \rightarrow_i b) \rightarrow_i^2 (c \rightarrow_i d)$.

We say that source i is trustworthy when meta-argument $trust(i)$ is acceptable, and we say that i provides evidence for argument a (of the attack $a \rightarrow b$) when $a \in Ar_i$ (when $a \rightarrow b \in \rightarrow_i$), and $trust(i) \longmapsto W_{acc(a)}$ ($trust(i) \longmapsto W_{Y_{a,b}}$).

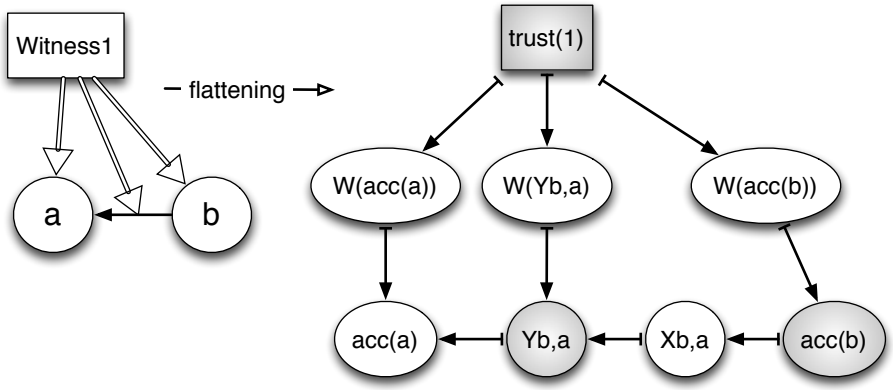


Figure 3.4: Introducing the sources in the argumentation frameworks.

Example 23. Consider the informal dialogue provided in the introduction. We represent the sources in the argumentation framework, as shown in Figure 3.4. Witness1 proposes a and b and the attack $a \rightarrow b$. Using the flattening function of Definition 46, we add meta-argument $trust(1)$ for representing Witness1 in the framework, and we add meta-arguments $acc(a)$ and $acc(b)$ for the arguments of Witness1. Witness1 provides evidence for these arguments, and the attack $b \rightarrow a$ by attacking the respective auxiliary arguments W. In the remainder of the chapter, we model the other conflicts highlighted in the dialogue.

Let $trust(i)$ be the information source i and $acc(a)$ and $Y_{a,b}$ the argument $a \in Ar_i$ and the attack $a \rightarrow b \in \rightarrow_i$ respectively, as defined in Definitions 43 and 44. Meta-argument $trust(i)$ can provide evidence for

$acc(a)$ and $Y_{a,b}$. Sources can attack other sources as well as their arguments and attacks. With a slight abuse of notation, we write $a \in \mathscr{E}'(EAF)$, even if the latter is a set of extensions, with the intended meaning that a is in some of the extensions of $\mathscr{E}'$. We now provide some properties of our model.

Proposition 16. If an argument $a \in Ar$ is not motivated by evidence, i.e., $a \notin Ar_i$ for all i, then a is not accepted, $a \notin \mathscr{E}'(EAF)$.

Proof. We assume admissibility based semantics. We prove the contrapositive: if argument a is accepted, then argument a is motivated by evidence. Assume argument a is accepted. Then auxiliary argument $W_{acc(a)}$ is rejected due to the conflict-free principle. Meta-argument $acc(a)$ is defended, so $W_{acc(a)}$ is attacked by an accepted argument using admissible semantics. Auxiliary argument $W_{acc(a)}$ can only be attacked by meta-argument $trust(i)$. We conclude that a is motivated by evidence. □

Proposition 16 is strengthened to Proposition 17.

Proposition 17. If there is no evidence for argument a, $a \notin Ar_i$, then the extensions $\mathscr{E}'(EAF)$ are precisely the same as the extensions of an EAF in which $a \notin Ar$, and where there are no attacks on or by a, i.e., there is no argument b attacking argument a and there is no argument c attacked by argument a.

Proof. We assume admissibility based semantics. Assume argument a is not motivated by evidence. This means that meta-argument $W_{acc(a)}$ is accepted, and meta-argument $acc(a)$ is rejected. Assume there exist an argument b such that b attacks a, $b \to a$, and an argument c such that a attacks c, $a \to c$. We prove that the extensions of the EAF with argument a are precisely the same as the extensions of the AF in which a does not exist, and there are no attacks on or by a. We use case analysis.

Case 1 Assume arguments b and c are not attacked, or they are attacked by unaccepted arguments. Then, we have that meta-argument $acc(b)$ is accepted and meta-argument $Y_{b,a}$ is accepted, and meta-argument $acc(c)$ is accepted, meta-argument $X_{a,c}$ is accepted, $Y_{a,c}$ is rejected, and $acc(a)$ is rejected due to the conflict-free principle because it is attacked by the accepted meta-argument $W_{acc(a)}$. The extension of this EAF includes b and c, but it does not include a.

Case 2 Assume arguments b and c are attacked by accepted arguments. Then, we have that meta-argument $acc(b)$ is rejected and meta-argument $Y_{b,a}$ is rejected, meta-argument $acc(c)$ is rejected, meta-argument $X_{a,c}$ is accepted, $Y_{a,c}$ is rejected, and $acc(a)$ is rejected due to the conflict-free principle. The extension of this EAF does not include a, b, and c.

Case 3 Assume argument b is not attacked or it is attacked by unaccepted arguments, and c is attacked by accepted arguments. Then, we have that meta-argument $acc(b)$ is accepted and meta-argument $Y_{b,a}$ is accepted, meta-argument $acc(c)$ is rejected, meta-argument $X_{a,c}$ is accepted, $Y_{a,c}$ is rejected, and $acc(a)$ is not accepted due to the conflict-free principle. The extension of this EAF includes b, but it does not include a and c.

Case 4 Assume argument b is attacked by accepted arguments and c is not attacked or it is attacked by unaccepted arguments. Then, we have that meta-argument $acc(b)$ is rejected and meta-argument $Y_{b,a}$ is rejected, meta-argument $acc(c)$ is accepted, meta-argument $X_{a,c}$ is accepted, $Y_{a,c}$ is rejected, and $acc(a)$ is rejected due to the conflict-free principle. The extension of this EAF includes c, but it does not include a and b.

Now we consider the same EAF without argument a, such that the attacks $b \rightarrow a$ and $a \rightarrow c$ do not exist.

Case 1 Assume arguments b and c are not attacked, or they are attacked by unaccepted arguments. Then, we have that meta-argument $acc(b)$ is accepted and meta-argument $acc(c)$ is accepted too. Each extension of this AF includes b and c.

Case 2 Assume arguments b and c are attacked by accepted arguments. Then, we have that meta-argument $acc(b)$ is rejected due to the conflict-free principle, and meta-argument $acc(c)$ is rejected too. Each extension of this AF does not include b, and c.

Case 3 Assume argument b is not attacked or it is attacked by unaccepted arguments, and c is attacked by accepted arguments. Then, we have that meta-argument $acc(b)$ is accepted, and meta-argument $acc(c)$ is rejected due to the conflict-free principle. Each extension of this AF includes b, but it does not include c.

Case 4 Assume argument b is attacked by accepted arguments and c is not attacked or it is attacked by unaccepted arguments. Then, we have that meta-argument $acc(b)$ is rejected due to the conflict-free principle, and meta-argument $acc(c)$ is accepted. Each extension of this AF includes c, but it does not include b.

Thus, the extensions of the EAF including argument a without evidence, and the EAF not including argument a are the same. □

Proposition 18. If there is no evidence for attack $a \rightarrow b$, i.e., $a \rightarrow b \not\in \rightarrow_i$, then the extensions $\mathscr{E}'(EAF)$ are precisely the same as the extensions of the EAF, in which the attack does not exist, $a \rightarrow b \not\in \rightarrow$.

The proof of Proposition 18 follows the proof of Proposition 17.

Proposition 19. Assume EAF is a framework in which argument a is motivated by evidence by the trustworthy source i, and there is another trustworthy source j. In that case, the extensions are the same if also j provides an evidence for a.

Proof. We assume admissibility based semantics. Assume argument a is motivated by evidence by the trustworthy source i. This means that $trust(i)$ is accepted. It provides evidence for argument a which means that meta-argument $trust(i)$ attacks meta-argument $W_{acc(a)}$: meta-argument $trust(i)$ is accepted, thus meta-argument $X_{trust(i),W_{acc(a)}}$ is rejected, meta-argument $Y_{trust(i),W_{acc(a)}}$ is accepted and meta-argument $W_{acc(a)}$ is rejected. Thus, meta-argument $acc(a)$ is accepted. We use case analysis.

Case 1 : Let argument a not be attacked or be attacked by unaccepted arguments. This means that meta-argument $acc(a)$ is accepted, and argument a is part of each extension of the EAF.

Case 2 : Let argument a be attacked by accepted arguments. This means that meta-argument $acc(a)$ is rejected, and argument a is not part of the extensions of the EAF.

Assume there is another trustworthy source j. This means that meta-argument $trust(j)$ is accepted. This source provides evidence for argument a, too. This means that $trust(j)$ attacks meta-argument $W_{acc(a)}$: meta-argument $trust(j)$ is accepted, thus meta-argument $X_{trust(j),W_{acc(a)}}$ is rejected, meta-argument $Y_{trust(j),W_{acc(a)}}$ is accepted and meta-argument $W_{acc(a)}$ is rejected. Thus, meta-argument $acc(a)$ is accepted. We use case analysis.

Case 1 : Let argument a not be attacked or be attacked by unaccepted arguments. This means that meta-argument $acc(a)$ is accepted, and argument a is part of each extension of the EAF.

Case 2 : Let argument a be attacked by accepted arguments. This means that meta-argument $acc(a)$ is rejected, and argument a is not part of the extensions of the EAF.

Thus, the extensions of the EAF are the same if there is also another source j, in addition to i, motivating by evidence argument a. □

Evidence for arguments. The evidence in favor of the arguments is an evidence provided by the agents for the arguments/attacks they propose. At the meta-level, this is modeled as an attack from meta-argument $trust(i)$ to W auxiliary arguments. However, there are other cases in which more evidence is necessary to motivate the acceptability of an argument. Consider the case of Witness1. His trustworthiness is attacked by Witness2. What happens to the evidence provided by Witness1? Since the source is not trustworthy then it cannot provide evidence. Meta-argument $trust(1)$ becomes rejected and the same happens to all its arguments and attacks. What is needed to make them acceptable again is more evidence. This evidence can be provided under the form of another argument which reinstates the acceptability of these information items.

Definition 46 allows only the sources to directly provide evidence for the information items. As for Witness5 and Witness6 in the dialogue, sources can provide evidence also by means of other arguments. This cannot be represented using the extended argumentation framework of Definition 46, this is why we need to extend it with an evidence relation $\looparrowright$ representing evidence provided under the form of arguments for the information items of the other sources.

Definition 47. A $TEAF^2$ with evidence is a tuple $\langle Ar, \langle Ar_1, \rightarrow_1, \rightarrow_1^2, \looparrowright_1 \rangle, \ldots, \langle Ar_n, \rightarrow_n, \rightarrow_n^2, \looparrowright_n \rangle, \rightarrow \rangle$ where $\looparrowright_i$ is a binary relation on $Ar_i \times Ar_j$ and the set of meta-arguments MA is $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in Ar_1 \cup \ldots \cup Ar_n\} \cup \{X_{a,b}, Y_{a,b} \mid a,b \in Ar_1 \cup \ldots \cup Ar_n\} \cup \{W_{acc(a)} \mid a \in Ar_1 \cup \ldots \cup Ar_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that hold the conditions of Definition 46, and: $acc(a) \longmapsto W_{acc(b)}$ iff $a \in Ar_i, b \in Ar_j$ and $a \looparrowright_i b$, and $W_{acc(b)} \longmapsto acc(b)$ iff $a \in Ar_i, b \in Ar_j$ and $a \looparrowright_i b$.

We say that a source i provides evidence in favor of the evidence provided by other source j to argument a when $a \in Ar_i, b \in Ar_j$, and $acc(a) \longmapsto W_{acc(b)}$.

The following properties hold for Definition 47.

Proposition 20. If there are multiple arguments $a_1 \in Ar_1, \ldots, a_n \in Ar_n$ motivating by evidence an argument $b \in Ar_k$ (or an attack), and there are no attacks on the arguments, $c_1 \rightarrow a_1 \notin \rightarrow_1, \ldots, c_n \rightarrow a_n \notin \rightarrow_n$, then b (or the attack) is accepted, $b \in \mathscr{E}'(EAF)$, iff at least one of the sources motivating by evidence the arguments $a_1, \ldots, a_n$ is trustworthy, i.e., $trust(j) \in \mathscr{E}(f(EAF))$ with $j \in 1, \ldots, n$.

Proof. Assume argument b is not directly motivated by evidence by an information source or the source supporting it is untrustworthy, and assume admissibility based semantics. This means that meta-argument $W_{acc(b)}$ is accepted, and meta-argument $acc(b)$ is rejected due to the conflict-free principle. Assume now that argument b is not attacked by other arguments, or it is attacked by unaccepted arguments, and assume there are n arguments $a_1, \ldots, a_n$ motivating by evidence argument b. Assume there not exist argument c_i such that it attacks a_i, and c_i is accepted.

First, we show that if there is at least one trustworthy source proposing an argument a_i which provides evidence for argument b, then b is accepted. This means that $trust(i)$ is accepted for $1 \leq i \geq n$. Then $W_{acc(a_i)}$ is rejected, and $acc(a)$ is accepted, $W_{acc(b)}$ is rejected and $acc(b)$ is accepted.

Now, we show that if argument b is accepted then there is at least one trustworthy source motivating it by evidence through argument a_i. This means that $acc(b)$ is accepted, and $W_{acc(b)}$ is not accepted. Thus there is at least on $Y_{acc(a_i),W_{acc(b)}}$ which is accepted. This means that $acc(a_i)$ is accepted, and $trust(i)$ is accepted. □

Proposition 21. Suppose two sources i and j provide evidence through arguments b and c respectively for the same argument a, i.e., $b \leadsto a \in \leadsto_i$ and $c \leadsto a \in \leadsto_j$, then it is the same whether a trustworthy source k provides evidence in favor of the evidence provided by i or j, i.e., $d \in Ar_k$.

Proof. Assume source k is trustworthy and assume admissibility based semantics. Source k provides evidence for argument d. Assume there are no other attacks on d. This means that meta-argument $trust(k)$ attacks meta-argument $W_{acc(d)}$ and $W_{acc(d)}$ attacks meta-argument $acc(d)$. The accepted meta-arguments are $acc(d)$ and $trust(k)$. We use case analysis.

Case 1 : Let the sources i and j be trustworthy, and let their arguments not be attacked by other arguments. This means that meta-arguments $trust(i)$ and $trust(j)$ are accepted, meta-arguments $acc(b)$ and $acc(c)$ are accepted and meta-argument $acc(a)$ is accepted. The evidence of source k through argument d consists in an attack from meta-argument $acc(d)$ to meta-argument $W_{acc(b)}$ or to meta-argument $W_{acc(c)}$. Both these meta-arguments are rejected because of the attacks from $trust(i)$ and $trust(j)$, respectively.

Case 2 : Let the sources i and j be untrustworthy, and let their arguments not be attacked by other arguments. This means that meta-arguments $trust(i)$ and $trust(j)$ are rejected. Thus, meta-arguments $acc(b)$ and $acc(c)$ are rejected. The evidence provided through argument d by source k consists in an attack from meta-argument $acc(d)$ to meta-argument $W_{acc(b)}$ or $W_{acc(c)}$. Independently on which meta-argument is attacked, this means that meta-argument $acc(b)$ or meta-argument $acc(c)$ is accepted, meta-argument $W_{acc(a)}$ is rejected, and meta-argument $acc(a)$ is accepted.

Case 3 : Let source i (or j) be trustworthy and source j (or i) be untrustworthy, and let their arguments not be attacked by other arguments. This means that meta-argument $trust(i)$ is accepted and meta-argument $trust(j)$ is rejected, meta-argument $W_{acc(b)}$ is accepted and meta-argument $W_{acc(c)}$ is rejected, meta-argument $acc(b)$ is accepted and meta-argument $acc(c)$ is rejected. Thus meta-argument $W_{acc(a)}$ is not accepted and meta-argument $acc(a)$ is accepted. The evidence provided through argument d to argument a does not change if meta-argument $acc(d)$ attacks meta-argument $W_{acc(b)}$ or $W_{acc(c)}$, because meta-argument $W_{acc(a)}$ is attacked by both $acc(b)$ and $acc(c)$. We conclude that argument a is accepted independently from the evidence provided by argument d.

□

Example 24. Consider the dialogue in the introduction. Argument g by Witness6 is an evidence for argument f by Witness5. This evidence is expressed in meta-argumentation in the same way as evidence provided by the sources, such as an attack to $W_{acc(f)}$ attacking $acc(f)$. In this case, it is meta-argument $acc(g)$ which attacks $W_{acc(f)}$, as visualized in Figure 3.5.

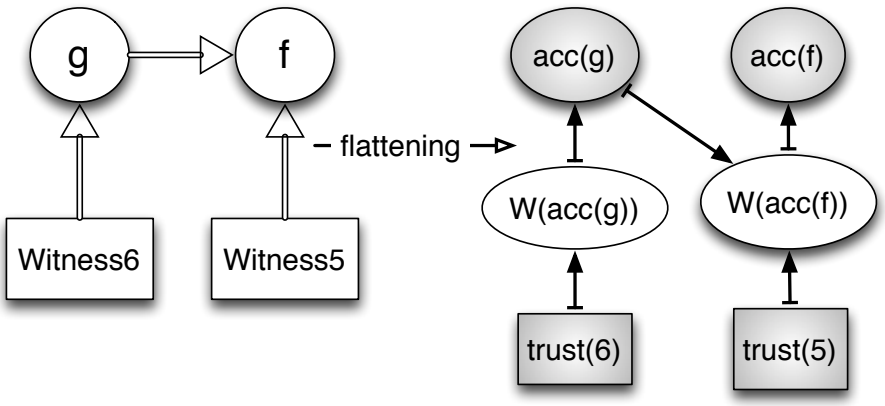


Figure 3.5: Introducing evidence for the arguments.

Focused trust relationships. In our model, trust is represented as the absence of an attack towards the sources or towards their information items, and as the presence of evidence in favor of the pieces of information. On the contrary, the distrust relationship is modeled as a lack of evidence in favor of the information items or as a direct attack towards the sources and their pieces of information.

In the informal dialogue, Witness2 attacks the trustworthiness of Witness1 as a credible witness. In this way, she is attacking each argument and attack proposed by Witness1. Witness4, instead, is not arguing against Witness3 but she is arguing against the attack $d \rightarrow b$ as it is proposed by Witness3. Finally, for Witness2 the untrustworthiness of Witness6 is related only to the argument g. We propose a focused view of trust in which the information sources may be attacked for being untrustworthy or for being untrustworthy only concerning a particular argument or attack. Definition 48 presents an EAF in which a new relation DT between sources is given to represent distrust.

Definition 48. A trust-based extended argumentation framework $DTEAF^2$ is a tuple $\langle Ar, \langle Ar_1, \rightarrow_1, \rightarrow_1^2, \not\rightarrow_1, DT_1\rangle, \ldots, \langle Ar_n, \rightarrow_n, \rightarrow_n^2, \not\rightarrow_n, DT_n\rangle, \rightarrow\rangle$ where for each source $1 \leq i \leq n$, $Ar_i \subseteq Ar \subseteq \mathscr{U}$ is a set of arguments, $\rightarrow \subseteq Ar \times Ar$, $\rightarrow_i \subseteq Ar_i \times Ar_i$ is a binary relation, $\rightarrow_i^2$ is a binary relation on $(Ar_i \cup \rightarrow_i) \times \rightarrow_i$, $\not\rightarrow_i$ is a binary relation on $Ar_i \times Ar_j$, and $DT \subseteq Ar_i \times \vartheta$ is a binary relation such that $\vartheta = j$ *or* $\vartheta \in Ar_j$ *or* $\vartheta \in \rightarrow_j$.

Definition 49 shows how to instantiate a $DTEAF^2$ enriched with a distrust relation with meta-arguments. In particular, the last three points of Definition 49 model, respectively, a distrust relationship towards an agent, a distrust relationship towards an argument, and a distrust relationship towards an attack. The unflattening function g and the acceptance function $\mathscr{E}'$ are defined as above.

Definition 49. Given a $DTEAF^2 = \langle Ar, \langle Ar_1, \rightarrow_1, \rightarrow_1^2, \not\rightarrow_1, DT_1\rangle, \ldots, \langle Ar_n, \rightarrow_n, \rightarrow_n^2, \not\rightarrow_n, DT_n\rangle, \rightarrow\rangle$, see Definition 48, the set of meta-arguments MA is $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in Ar_1 \cup \ldots \cup Ar_n\} \cup \{X_{a,b}, Y_{a,b} \mid a,b \in Ar_1 \cup \ldots \cup Ar_n\} \cup \{W_{acc(a)} \mid a \in Ar_1 \cup \ldots \cup Ar_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that hold the conditions of Definitions 46 and 47, and:

- $acc(a) \longmapsto X_{a,b}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $X_{a,b} \longmapsto Y_{a,b}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $Y_{a,b} \longmapsto acc(b)$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and
- $trust(i) \longmapsto X_{trust(i),W_{acc(a)}}$ iff $a \in Ar_i$, and $X_{trust(i),W_{acc(a)}} \longmapsto Y_{trust(i),W_{acc(a)}}$ iff $a \in Ar_i$, and $Y_{trust(i),W_{acc(a)}} \longmapsto W_{acc(a)}$ iff $a \in Ar_i$, and $W_{acc(a)} \longmapsto acc(a)$ iff $a \in Ar_i$, and

- $trust(i) \longmapsto X_{trust(i),W_{Y_{a,b}}}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $X_{trust(i),W_{Y_{a,b}}} \longmapsto Y_{trust(i),W_{Y_{a,b}}}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $Y_{trust(i),W_{Y_{a,b}}} \longmapsto W_{Y_{a,b}}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $W_{Y_{a,b}} \longmapsto Y_{a,b}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and

- $trust(i) \longmapsto W_{acc(a)}$ iff $a \in Ar_i$ and $aDT_i trust(j)$, and $W_{acc(a)} \longmapsto acc(a)$ iff $a \in Ar$ and $aDT_i trust(j)$, and $acc(a) \longmapsto X_{acc(a),trust(j)}$ iff $a \in Ar_i$ and $aDT_i trust(j)$, and $X_{acc(a),trust(j)} \longmapsto Y_{acc(a),trust(j)}$ iff $a \in Ar_i$ and $aDT_i trust(j)$, and $Y_{acc(a),trust(j)} \longmapsto trust(j)$ iff $a \in Ar_i$ and $aDT_i trust(j)$, and

- $trust(i) \longmapsto W_{acc(a)}$ iff $a \in Ar_i, b \in Ar_j$ and $aDT_i b$, and $W_{acc(a)} \longmapsto acc(a)$ iff $a \in Ar, b \in Ar_j$ and $aDT_i b$, and $acc(a) \longmapsto X_{acc(a),Y_{trust(j),W_{acc(b)}}}$ iff $a \in Ar_i, b \in Ar_j$ and $aDT_i b$, and $X_{acc(a),Y_{trust(j),W_{acc(b)}}} \longmapsto Y_{acc(a),Y_{trust(j),W_{acc(b)}}}$ iff $a \in Ar_i, b \in Ar_j$ and $aDT_i b$, and $Y_{acc(a),Y_{trust(j),W_{acc(b)}}} \longmapsto Y_{trust(j),W_{acc(b)}}$ iff $a \in Ar_i, b \in Ar_j$ and $aDT_i b$, and

- $trust(i) \longmapsto W_{acc(a)}$ iff $a \in Ar_i, b,c \in Ar_j$ and $aDT_i(b \rightarrow_j c)$, and $W_{acc(a)} \longmapsto acc(a)$ iff $a \in Ar, b,c \in Ar_j$ and $aDT_i(b \rightarrow_j c)$, and $acc(a) \longmapsto X_{acc(a),Y_{trust(j),W_{Y_{b,c}}}}$ iff $a \in Ar_i, b,c \in Ar_j$ and $aDT_i(b \rightarrow_j c)$, and $X_{acc(a),Y_{trust(j),W_{Y_{b,c}}}} \longmapsto Y_{acc(a),Y_{trust(j),W_{Y_{b,c}}}}$ iff $a \in Ar_i, b,c \in Ar_j$ and $aDT_i(b \rightarrow_j c)$, and $Y_{acc(a),Y_{trust(j),W_{Y_{b,c}}}} \longmapsto Y_{trust(j),W_{Y_{b,c}}}$ iff $a \in Ar_i, b,c \in Ar_j$ and $aDT_i(b \rightarrow_j c)$.

We say that a source j is untrustworthy when there is an attack from an argument $a \in Ar_i$ to j, $aDT_i trust(j)$. We say that an argument $a \in Ar_j$ or attack $a \rightarrow_j b \in \rightarrow_j$ is untrustworthy when there is an attack from an argument $c \in Ar_i$ to a or $a \rightarrow_j b$, $cDT_i a$ or $cDT_i(a \rightarrow_j b)$.

Proposition 22. Assume that source i is the only source motivating by evidence argument $a \in Ar_i$ and attack $c \rightarrow b \in \rightarrow_i$. If the information source i is considered to be untrustworthy, then a and $c \rightarrow b$ are not acceptable.

Proof. We assume admissibility based semantics. We prove the contrapositive: if the arguments and attacks supported by an information source i are acceptable then the information source i is considered to be trustworthy. Assume the source supports argument a and the attack $c \rightarrow b$ and assume that this argument and this attack are acceptable. Then auxiliary arguments $W_{acc(a)}$ and $W_{Y_{c,b}}$ are rejected due to the conflict-free principle. Meta-arguments $acc(a)$ and $Y_{c,b}$ are defended, thus $W_{acc(a)}$ and $W_{Y_{c,b}}$ are attacked by an acceptable argument, using admissible semantics. We assumed that this argument and this attack have no other evidence, so auxiliary arguments $W_{acc(a)}$ and $W_{Y_{c,b}}$ can only be attacked by meta-argument $trust(i)$. Since they are attacked by an acceptable argument, we conclude that the source i is acceptable. □

Example 25. Figure 3.6.a shows that Witness2 attacks the trustworthiness of Witness1 by means of argument c. In meta-argumentation, we have that $trust(2)$ provides evidence for $acc(c)$ by attacking meta-argument $W_{acc(c)}$ and, with meta-arguments X,Y, it attacks $trust(1)$. This means that if Witness1 is untrustworthy then each of his arguments and attacks cannot be acceptable either, if there is no more evidence. The set of acceptable arguments for the meta-argumentation framework is $\mathscr{E}(f(focus1)) = \{trust(2), acc(c), Y_{acc(c),trust(1)}\}$. In Figure 3.6.b-c, instead, the attack is directed against a precise information item provided by the source. In particular, Witness4 attacks the attack $d \rightarrow b$ as provided by Witness3. This is achieved in meta-argumentation by means of an attack from meta-argument $acc(e)$, for which $trust(4)$ provides evidence, to the attack characterized by auxiliary argument $Y_{d,b}$. The set of acceptable arguments is $\mathscr{E}(f(focus2)) = \{trust(4), trust(3), acc(d), acc(e), acc(b), Y_{acc(e),Y_{trust(3),W_{Y_{b,d}}}}, W_{Y_{d,b}}\}$. Witness3's attack $d \rightarrow$

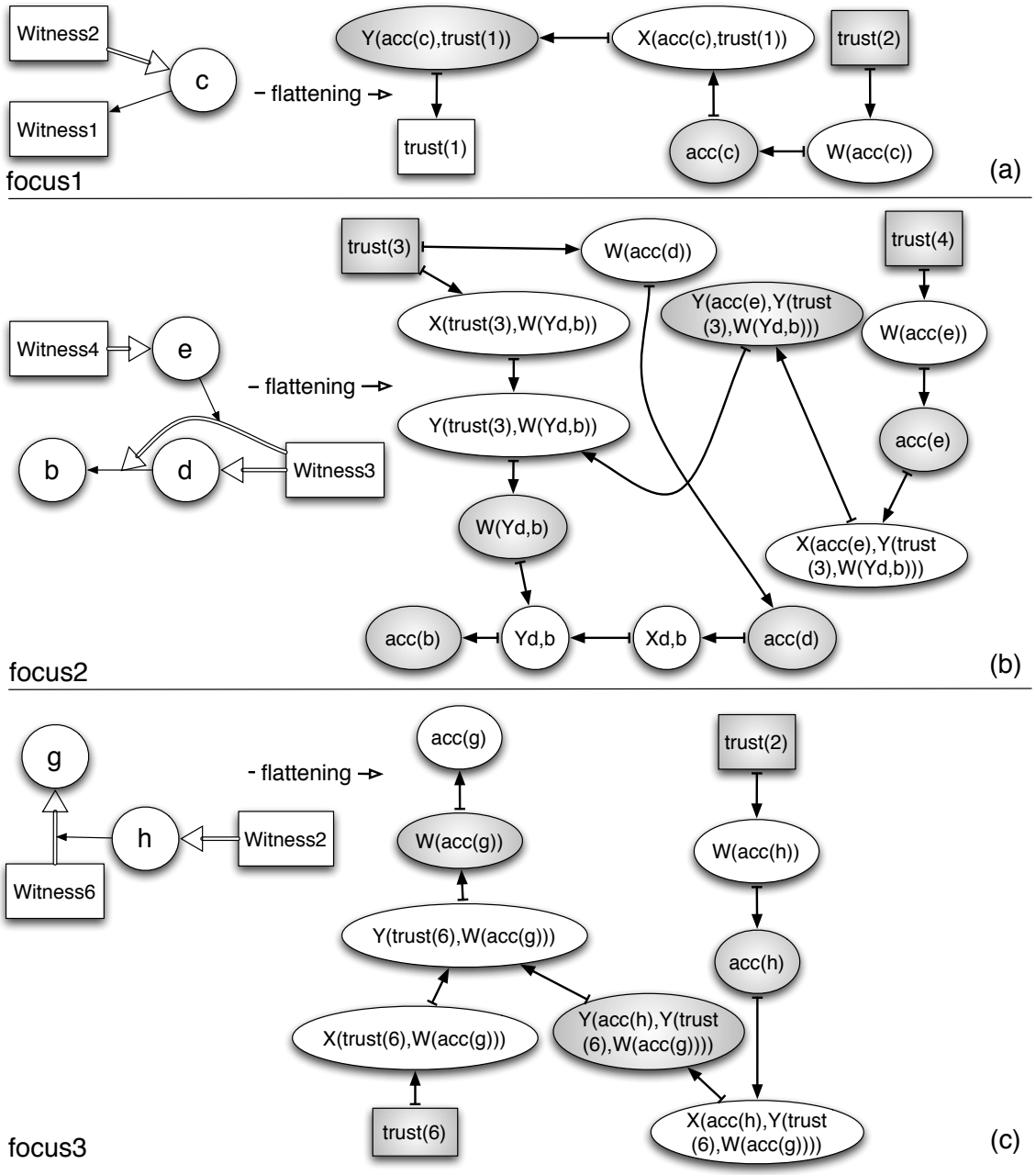


Figure 3.6: Focused trust in argumentation.

b is evaluated as untrustworthy by Witness4 and thus it is not acceptable. Finally, Witness2 evaluates Witness6 as untrustworthy concerning argument g. In meta-argumentation, $trust(2)$, by means of meta-argument $acc(h)$, attacks meta-argument $acc(g)$ proposed by $trust(6)$. The set of acceptable arguments is $\mathscr{E}^{\circ}(f(focus3)) = \{trust(2), trust(6), acc(h), Y_{acc(h),Y_{trust(6),W_{acc(g)}}}, W_{acc(g)}\}$.

Feedback from information items to sources and back. In the previous sections, we have introduced the information sources in the argumentation framework in order to deal with the conflicts about trust. Moreover, in our framework, the agents are allowed to attack the trustworthiness of the other information sources or the trustworthiness of the single information items the sources propose. The relation, concerning trust, among the sources and the arguments or attacks they motivate is in one direction only. In particular, if an agent is considered to be untrustworthy, then also all the information items proposed by such an agent are considered untrustworthy. But what happens to the trustworthiness of an agent which is not directly attacked but it has all its information items (or at least *n* information items) attacked? In the current framework, these attacked items do not effect the trustworthiness of the sources proposing them, e.g., if a source has the trustworthiness of all its information items attacked, the source's trustworthiness is accepted.

The idea proposed by Castelfranchi and Falcone [90] is that there is a bidirectional link between the source and its information items: the provided data is more or less believable on the basis of the source's trustworthiness, but there is feedback such that the invalidation of the data feeds back on the sources' credibility. The overall amount and sign (increment or decrement) of the feedback depends on how much the overall quality of the message surprises the agent, with respect to its prior assessment of the source trustworthiness. This captures the principle that information quality should change one's assessment of its source only when the agent learns something new about the capacity of the source to deliver information of either high or low quality. In other words, there should be a feedback on the source only when the quality of its argument tells me something new about the source's trustworthiness, revealing my previous opinion to be wrong. Otherwise, the quality of the new argument just confirms my previous assessment of the source, and confirmation, by definition, consolidates a pre-existing judgment, rather than modifying it. This points to the role of prediction in feedback dynamics from arguments to sources, and this prediction is based on the pre-existing degree of trustworthiness of the source of a given argument. In this chapter, we do not represent the increment of the feedback towards the information source. In our framework, a trustworthy source is mirrored in an accepted meta-argument of the kind $trust(i)$, and this acceptability cannot be improved. The representation of this kind of feedback would be possible in numerical approaches to trust representation in argumentation, as proposed for instance by da Costa Pereira et al. [113] and Parsons et al. [253].

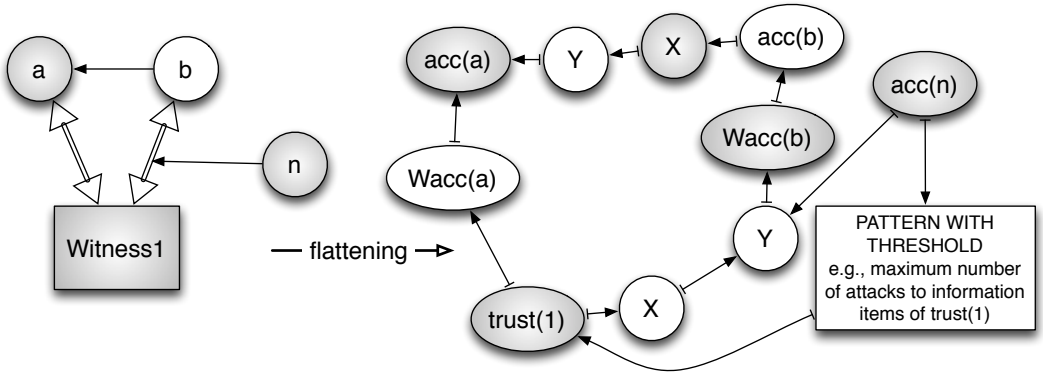


Figure 3.7: Feedback between the information items and sources.

In this section, we rely on this analysis of the trust dynamics phenomenon, in order to model the feedback from the information items to the sources. For instance, the fact that the major part of the arguments of a source are considered untrustworthy is seen as a negative experience, and leads to the decrease of the

trustworthiness of the sources itself. In this chapter, we do not consider the unpredictable cases analyzed by Castelfranchi and Falcone [90], where trust decreases with positive experiences, and increases with negative ones. The representation of these cases is left as future work.

We introduce the feedback from the information items to the sources, in such a way that, following different criteria, the untrustworthiness of the items influences the trustworthiness of its source. The general idea of our approach is visualized in Figure 3.7. First, we insert in the framework a pattern which is activated if the number of attacks to this pattern exceeds a certain threshold. In this case, the pattern activates an attack towards the meta-argument representing the information source. The activation pattern is visualized in Figure 3.8, and it comes from the idea of conjunctive and proof standard patterns defined by Villata et al. [308]. The arguments attacking the information items proposed by the source attack also the pattern, in particular, each argument *arg* attacking the items attacks also one of the X meta-arguments of the pattern. These meta-arguments conjunctively attack argument s, which attacks the meta-argument representing the source. The pattern acts like a filter that raises the attack against the source only if the amount of incoming attacks is achieved.

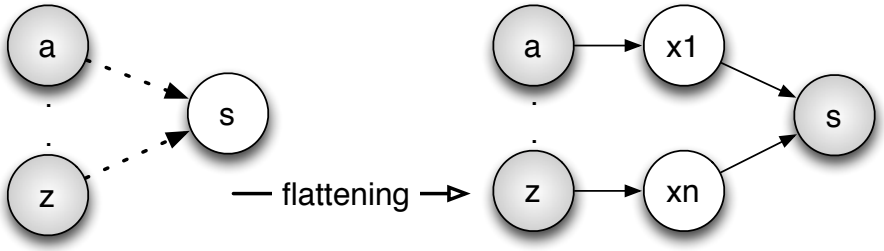


Figure 3.8: The activation pattern with a threshold of n arguments.

Second, for each attack to the trustworthiness of one of the information items of a source, this attack is duplicated and it is addressed also towards the pattern which attacks the information source. Summarizing, every attack to the arguments or attacks of a source is addressed also towards the pattern which has the aim to attack directly the trustworthiness of the source, if the number of attack exceeds the given threshold.

Example 26. Let us consider now the example proposed in Figure 3.7. In the informal dialogue, Witness1 proposes two arguments a and b, and the attack between them. Consider now the introduction of a new argument n, which attacks the trustworthiness of argument b as proposed by Witness1. In the flattened framework, the meta-argument $trust(1)$ provides evidence for meta-arguments $acc(a)$ and $acc(b)$ by attacking the auxiliary arguments $W_{acc(a)}$ and $W_{acc(b)}$. The attack of the new argument is addressed from meta-argument $acc(n)$ to the auxiliary argument $Y_{W_{acc(b)}}$ which attacks $W_{acc(b)}$. Since we are interested in modeling also the feedback from the information items to the sources, we add an additional attack from meta-argument $acc(n)$ to the pattern we use to measure the number of attacks to the information items proposed by Witness1. From this pattern, an attack is raised against the meta-argument $trust(1)$. If the number of attacks towards the pattern overcomes the given threshold, then the attack against $trust(1)$ becomes active, and $trust(1)$ becomes unacceptable, i.e., Witness1 is considered untrustworthy: so argument a is not acceptable.

We model feedback using the pattern associated with the threshold in order to maintain the choice of meta-argumentation, and avoiding the introduction of numerical techniques, as done for instance by da Costa

Pereira et al. [113].

Modeling trust as a multidimensional concept. In this section, we investigate two dimensions of trust that have to be independently evaluated such as the sincerity or credibility of a source and its competence. We simplify Castelfranchi and Falcone's model [90], and focus only on two broad categories of relevant features in the source: competence (to what extent the source is deemed able to deliver a correct argument), and sincerity (to what extent the source is considered willing to provide a correct argument), both of which contribute to determine the source's overall trustworthiness. The evaluations of competence and sincerity are allowed to change across different domains. For instance, a reliable doctor will be considered competent in the health domain, but not necessarily so when suggesting a restaurant; conversely, a food critic is typically assumed to be trustworthy on the latter domain but not on the former. Similarly, one might think that a colleague who is competing with her for a promotion is likely to be insincere in giving her tips on how to improve her career, and yet there is no reason to doubt his sincerity when he suggests a movie. Here, we consider competence and sincerity as two possible dimensions for assessing the trustworthiness of a source.

We represent competence and sincerity using meta-arguments, and the attacks to these meta-arguments represent the conflicts about trust regarding a precise dimension of trust. The introduction in our framework of these two dimensions is visualized in Figure 3.9. We start from the usual situation in which an information source supports an argument, namely Witness7 supports argument i in the informal dialogue. We want to distinguish the two possible conflicts concerning argument i: a conflict meaning that Witness7 is considered untrustworthy on the competence regarding argument i, and a conflict meaning that Witness7 is considered untrustworthy on the sincerity in proposing argument i. An example of the first case is given in the dialogue by the attack of argument l to argument i, and an example of the second case is given by the attack of argument m to argument i. Note that even if both arguments l and m attack argument i, they attack different dimensions of argument i.

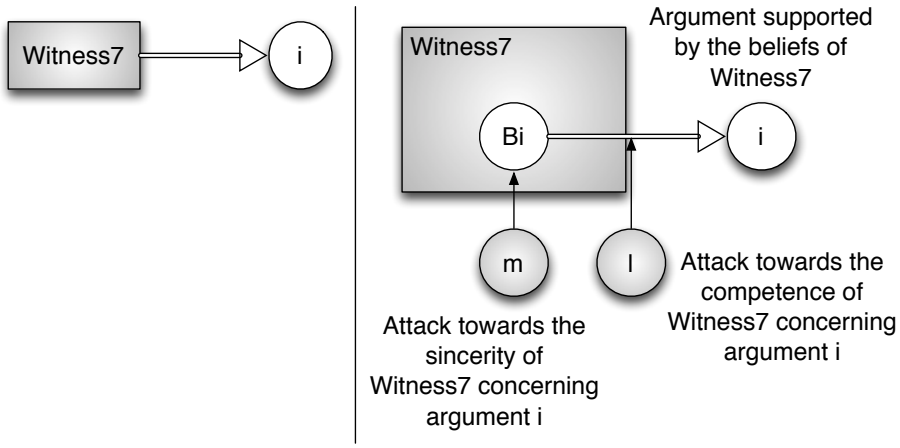


Figure 3.9: Modelling competence and sincerity.

We model sincerity and competence as visualized in Figure 3.9. Meta-argument B_i represents the belief associated to the information source concerning argument i, and it means "*the source believes argument i*" where argument i is the argument supported by the beliefs of the source. The meta-argument B_i provides evidence in favor of argument i, as a result of the competence attributed to the source. In this framework,

an attack towards the sincerity of the source is addressed against the meta-argument representing the belief of the source, i.e., against meta-argument B_i. An attack towards the competence of the source is addressed, instead, against the evidence provided by meta-argument B_i to argument i. This attack means that the source believes argument i but it is not evaluated competent concerning i. Note that an attack towards argument i is treated as in the previous sections, since it is a direct attack towards the content of argument i.

Definition 50. A trust-based extended argumentation framework $DTEAF^2_{CS}$ is a tuple $\langle Ar, \langle Ar_1, \rightarrow_1, \rightarrow^2_1, \not\rightarrow_1, DT_1, DT_{1s}, DT_{1c}\rangle, \ldots, \langle Ar_n, \rightarrow_n, \rightarrow^2_n, \not\rightarrow_n, DT_n, DT_{ns}, DT_{nc}\rangle, \rightarrow\rangle$ where for each source $1 \leq i \leq n$, $Ar_i \subseteq Ar \subseteq \mathscr{U}$ is a set of arguments, $\rightarrow \subseteq Ar \times Ar$, $\rightarrow_i \subseteq Ar_i \times Ar_i$ is a binary relation, $\rightarrow^2_i$ is a binary relation on $(Ar_i \cup \rightarrow_i) \times \rightarrow_i$, $\not\rightarrow_i$ is a binary relation on $Ar_i \times Ar_j$, and $DT \subseteq Ar_i \times \vartheta$ is a binary relation such that $\vartheta = j$, and $DT_s \subseteq Ar_i \times \vartheta$ is a binary relation such that $\vartheta \in Ar_j$ *or* $\vartheta \in \rightarrow_j$, and $DT_c \subseteq Ar_i \times \vartheta$ is a binary relation such that $\vartheta \in Ar_j$ *or* $\vartheta \in \rightarrow_j$.

Definition 51 shows how to instantiate an extended argumentation framework enriched with a distrust relation, which distinguishes distrust concerning competence and sincerity. The unflattening function g and the acceptance function $\mathscr{E}'$ are defined as above.

Definition 51. Given a $DTEAF^2_{CS} = \langle Ar, \langle Ar_1, \rightarrow_1, \rightarrow^2_1, \not\rightarrow_1, DT_1, DT_{1s}, DT_{1c}\rangle, \ldots, \langle Ar_n, \rightarrow_n, \rightarrow^2_n, \not\rightarrow_n, DT_n, DT_{ns}, DT_{nc}\rangle, \rightarrow\rangle$, see Definition 50, the set of meta-arguments MA is $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in Ar_1 \cup \ldots \cup Ar_n\} \cup \{X_{a,b}, Y_{a,b} \mid a,b \in Ar_1 \cup \ldots \cup Ar_n\} \cup \{W_{acc(a)} \mid a \in Ar_1 \cup \ldots \cup Ar_n\} \cup \{B_a \mid a \in Ar_1 \cup \ldots \cup Ar_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that hold the conditions of Definitions 46, 47, and 49, and:

- $B_a \longmapsto X_{B_a,a}$ iff $a \in Ar_i$, and $X_{B_a,a} \longmapsto Y_{B_a,a}$ iff $a \in Ar_i$, and $Y_{B_a,a} \longmapsto W_{B_a,a}$ iff $a \in Ar_i$, and $W_{B_a,a} \longmapsto acc(a)$ iff $a \in Ar_i$ and
- $B_{a\rightarrow b} \longmapsto X_{B_{a\rightarrow b},a\rightarrow b}$ iff $a \rightarrow b \in \rightarrow_i$, and $X_{B_{a\rightarrow b},a\rightarrow b} \longmapsto Y_{B_{a\rightarrow b},a\rightarrow b}$ iff $a \rightarrow b \in \rightarrow_i$, and $Y_{B_{a\rightarrow b},a\rightarrow b} \longmapsto W_{B_{a\rightarrow b},a\rightarrow b}$ iff $a \rightarrow b \in \rightarrow_i$, and $W_{B_{a\rightarrow b},a\rightarrow b} \longmapsto Y_{a,b}$ iff $a \rightarrow b \in \rightarrow_i$ and
- $trust(i) \longmapsto X_{trust(i),W_{acc(a)}}$ iff $a \in Ar_i$, and $X_{trust(i),W_{acc(a)}} \longmapsto Y_{trust(i),W_{acc(a)}}$ iff $a \in Ar_i$, and $Y_{trust(i),W_{acc(a)}} \longmapsto W_{acc(a)}$ iff $a \in Ar_i$, and $W_{acc(a)} \longmapsto B_a$ iff $a \in Ar_i$, and
- $trust(i) \longmapsto X_{trust(i),W_{Y_{a,b}}}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $X_{trust(i),W_{Y_{a,b}}} \longmapsto Y_{trust(i),W_{Y_{a,b}}}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $Y_{trust(i),W_{Y_{a,b}}} \longmapsto W_{Y_{a,b}}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and $W_{Y_{a,b}} \longmapsto B_{a\rightarrow b}$ iff $a,b \in Ar_i$ and $a \rightarrow_i b$, and
- $acc(a) \longmapsto B_b$ iff $a \in Ar_i, b \in Ar_j$ and $aDT_{is}b$, and
- $acc(a) \longmapsto Y_{B_a,a}$ iff $a \in Ar_i, b \in Ar_j$ and $aDT_{ic}b$, and
- $acc(a) \longmapsto B_{b\rightarrow c}$ iff $a \in Ar_i, b,c \in Ar_j$ and $aDT_{is}(b \rightarrow_j c)$, and
- $acc(a) \longmapsto Y_{B_{b\rightarrow c},b\rightarrow c}$ iff $a \in Ar_i, b,c \in Ar_j$ and $aDT_{ic}(b \rightarrow_j c)$.

We say that an argument $a \in Ar_i$ or attack $a \rightarrow b \in \rightarrow_i$ is untrustworthy concerning sincerity when there is an attack from an argument $c \in Ar_j$ to B_a or $B_{a\rightarrow b}$, $cDT_{js}a$ or $cDT_{js}(a \rightarrow b)$. We say that an argument $a \in Ar_i$ or attack $a \rightarrow b \in \rightarrow_i$ is untrustworthy concerning competence when there is an attack from an argument $c \in Ar_j$ to $Y_{B_a,a}$ or $Y_{B_{a\rightarrow b},a\rightarrow b}$, $cDT_{jc}a$ or $cDT_{jc}(a \rightarrow b)$.

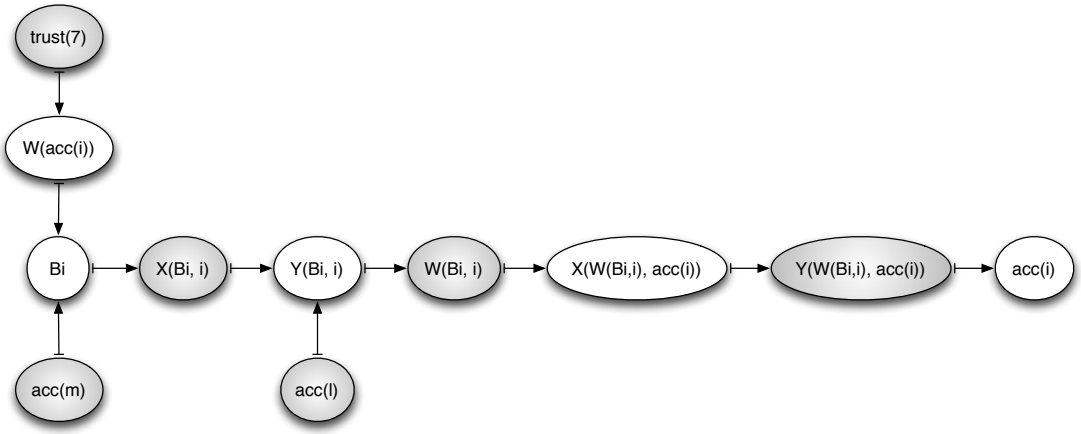


Figure 3.10: The flattening of the competence and sincerity's framework.

The flattening of the new framework distinguishing between attacks towards the sincerity of a source in proposing an information item, and attacks towards the competence of a source in proposing an information item is formalized in Definition 51. An example of flattening is visualized in Figure 3.10.

Example 27. The meta-argument representing Witness7, $trust(7)$, provides evidence by means of the auxiliary argument $W_{acc(i)}$ to meta-argument B_i representing the fact that argument i is believed by Witness7. If this meta-argument is accepted, it means that there are no doubts about the sincerity of Witness7 concerning argument i. Meta-argument B_i provides evidence for meta-argument $acc(i)$, representing argument i in the meta-level. This evidence is built in the same way as the evidence provided by the sources for their information items, which means that meta-argument B_i attacks, towards auxiliary arguments X and Y, the auxiliary argument $W_{B_i,i}$. This auxiliary argument attacks, always by means of X and Y auxiliary arguments, the meta-argument $acc(i)$. In this framework, the acceptability of meta-argument $acc(i)$ depends on the acceptability of the belief regarding argument i. An attack towards the competence of argument i, instead, is addressed against meta-argument $Y_{B_i,i}$. In this way, argument $acc(i)$ can be made unacceptable in two ways: (1) by attacking directly meta-argument B_i (sincerity), and (2) by attacking the attack from B_i to $W_{B_i,i}$ (competence). Figure 3.10 shows these two cases with the attacks from argument m and argument l, respectively.

The following property holds for our model of competence and sincerity.

Proposition 23. Suppose a trustworthy source i provides evidence for argument a, and another trustworthy source j provides evidence for an argument b where b attacks the trustworthiness of argument a, then the extensions are the same if argument b attacks the sincerity or the competence about argument a.

Proof. We assume admissibility based semantics. Assume argument a and argument b are not attacked by other arguments. We use case analysis.

Case 1 Let argument b attack the sincerity dimension of the trustworthiness of argument a. Meta-argument $acc(b)$ is accepted, due to the conflict-free principle, as motivated by evidence by a trustworthy source and not attacked by external arguments, and meta-argument $Y_{acc(b),B_a}$ is accepted. This means that meta-argument B_a is rejected, and thus argument $W_{B_a,a}$ is accepted, and meta-argument $acc(a)$ is

rejected. Argument a is not part of any admissibility-based extension, and argument b is part of the admissibility-based extensions.

Case 2 Let argument b attack the competence dimension of the trustworthiness of argument a. Meta-argument $acc(b)$ is accepted, as supported by a trustworthy source and not attacked by external arguments, and argument B_a is accepted, as supported by a trustworthy source. Then meta-argument $Y_{B_a,a}$ is rejected, as attacked by argument $acc(b)$, due to the conflict-free principle. This means that meta-argument $W_{B_a,a}$ is accepted, and meta-argument $acc(a)$ is rejected. Argument a is not part of any admissibility-based extension, and argument b is part of the admissibility-based extensions.

We conclude that the extensions are the same whether argument b attacks the sincerity or the competence of argument a. □

In the next section, we will highlight that in certain scenarios a numerical approach to reason on trust using argumentation theory is required, and we will introduce our fuzzy labeling algorithm to weight arguments in the framework depending on the trustworthiness degree of the source proposing them.

3.3 Fuzzy argumentation labeling for trust

In a multiagent environment, belief revision aims at describing the changes in the agent's mind in response to new information. On the other hand, one of the important concerns in argumentation is the strategies employed by an agent in order to succeed in changing the mind of another agent. To this aim, the agent must provide *good enough* reasons to (justify and then) succeed in such request of change. We can then view argumentation as an "incitement" to make an agent change its mind.

This section is not "just" about integrating belief revision and argumentation in a single framework. It aims at using the strength resulting from such a combination to solve the problem of loss of information in the case of reinstatement of previous information in multiagent systems. More precisely, we answer the question "in case of such a reinstatement, *how* to recover from the loss of previous information which should become acceptable with new information, and to which extent old information should be recovered?"

The proposed framework integrates the first three basic steps considered by Falappa and colleagues. Indeed, in order to represent real situations more faithfully, we consider that new information is associated with a degree of plausibility which represents the trustworthiness, for the agent, of the source of information. This is in line with some work in the literature, like, for example, [111], but the originality, which is the main difference with the previously cited authors, lies in the fact that a piece of information is represented as an argument which can be more or less acceptable. Therefore, such a degree directly influences the evaluation, performed by an agent, of new information and, as a consequence, it also influences the extent to which an agent changes its mind. Based on these considerations, we propose a fuzzy reinstatement algorithm which provides a satisfactory answer to our research question, which may be broken down into the following subquestions:

- How to represent arguments and beliefs in this setting?
- How to define a fuzzy evaluation of the arguments?
- How to address the change in the agent's cognitive state?

The first step is about determining the most suitable representation of partially trusted arguments and beliefs. Arguments, of the form $\langle \Phi, \phi \rangle$, support the agents' beliefs, which can be represented as the conclusions of structured arguments. The trustworthiness of a source can be measured by using probabilities only in the case in which data are available based on past experiences, for example. In many realistic cases, such data is not available. It is well known that possibilistic logic is well suited to deal with incomplete information. For example, [3] introduce a unified negotiation framework based on possibilistic logic to represent the agent's beliefs, preferences, decision, and revision under an argumentation point of view. A fuzzy labeling will then determine the fuzzy set of the agent's beliefs. [111] adopt the representation of uncertain beliefs proposed in [130]. The main point of their proposal may be described as belonging to the fourth among the basic steps proposed by [143], in the sense that they derive the most useful goals from the most plausible beliefs and desires. However, their approach for representing the changes in the agent's beliefs is not argumentation-based and cannot treat reinstatement in a satisfactory way.

The second step is about defining an algorithm which allows a fuzzy evaluation of the arguments. In crisp argumentation, arguments are evaluated, following a specific semantics, as acceptable or not acceptable, as shown by [133]. Intuitively, accepted arguments are those arguments which are not attacked by other accepted arguments and unaccepted arguments are those attacked by accepted arguments. Given an accepted argument, its conclusion can be adopted as belief in the agent's belief base. To represent the degrees of trust, we rethink the usual crisp argument evaluation [133, 84] by evaluating arguments in terms of fuzzy degrees of acceptability.

The third step is the choice about how to address the change in the cognitive state of the agent. As observed by [129] and [123], for example, the main approaches to belief revision adopt the principle of the "priority to incoming information" but, in the context of multiagent systems, this principle presents some drawbacks. In particular, in a static situation, the chronological sequence of arrival of distinct pieces of information has nothing to do with their trustability or importance. This is supported also by [151], where revision algorithms are presented in order to take into account the history of previous revisions as well as possible revision options which were first discarded but may now be pursued. The assumption we put forward in this chapter is that, even if the agent accepts the incoming information throwing away part of the previously adopted belief base, this change must not be irrevocable. This means that, in the future, new information may turn the tide in such a way to have the past incoming information excluded from the belief-base and the original belief *somehow* reintegrated. This is exactly what happens in argumentation under the name of *reinstatement* principle. The difference, which is also one of the original contributions of this chapter, is that the extent of the integration depends on the agent's trust in the source. Indeed, we evaluate arguments in a gradual way depending on such a degree of trust.

A schematic illustration of the proposed framework is visualized in Figure 3.11. The framework may be regarded as a belief revision model, based on argumentation. An agent interacts with the world by receiving arguments A from one or more *sources*. The agent's internal mental state is completely described by a fuzzy set of trustful arguments Ar, from which the beliefs of the agent may be derived. A *trust* module, whose details are not covered in this chapter, assigns a trust degree τ to each source. As new arguments A are received, they are added to Ar with the same membership degree as the degree τ to which their source is trusted. Fuzzy labeling of Ar yields a fuzzy reinstatement labeling α, which may be regarded as a fuzzy set of acceptable arguments, whose consequences induce a possibility distribution π, from which an explicit representation $\mathbf{B}$ of the agent's beliefs is constructed as the necessity measure N of possibility distribution π. Notice that we do not make any further assumptions on the trust model. This is out of the scope of this chapter.

A classical propositional language may be used to represent information for manipulation by a cognitive

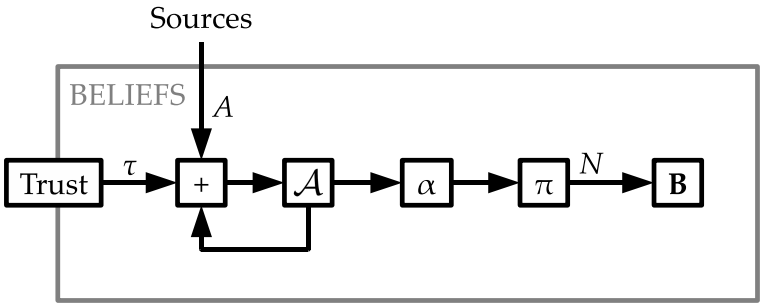


Figure 3.11: A schematic illustration of the proposed framework.

agent.

Definition 52 (Language)**.** Let Prop be a *finite*[1] set of atomic propositions and let $\mathscr{L}$ be the propositional language such that $\text{Prop} \cup \{\top, \bot\} \subseteq \mathscr{L}$, and, $\forall \phi, \psi \in \mathscr{L}$, $\neg\phi \in \mathscr{L}$, $\phi \wedge \psi \in \mathscr{L}$, $\phi \vee \psi \in \mathscr{L}$.

As usual, one may define additional logical connectives and consider them as useful shorthands for combinations of connectives of $\mathscr{L}$, e.g., $\phi \supset \psi \equiv \neg\phi \vee \psi$.

We will denote by $\Omega = \{0,1\}^{\text{Prop}}$ the set of all interpretations on Prop. An interpretation $\mathscr{I} \in \Omega$ is a function $\mathscr{I} : \text{Prop} \to \{0,1\}$ assigning a truth value $p^{\mathscr{I}}$ to every atomic proposition $p \in \text{Prop}$ and, by extension, a truth value $\phi^{\mathscr{I}}$ to all formulas $\phi \in \mathscr{L}$. We will denote by $[\phi]$ the set of all models of ϕ, $[\phi] = \{\mathscr{I} : \mathscr{I} \models \phi\}$.

We can give the arguments a structure, and the attack relation is defined in terms of such a structure of the arguments, following the example of [41].

Definition 53. An argument is a pair $\langle \Phi, \phi \rangle$, with $\phi \in \mathscr{L}$ and $\Phi \subset \mathscr{L}$, such that

1. $\Phi \nvdash \bot$,
2. $\Phi \vdash \phi$, and
3. Φ is minimal w.r.t. set inclusion.

We say that $\langle \Phi, \phi \rangle$ is an argument for ϕ. We call ϕ the conclusion and Φ the support of the argument.

The more specific forms of conflict are called *undercut* and *rebuttal*.

Definition 54. (Undercut, Rebuttal) An *undercut* for an argument $\langle \Phi, \phi \rangle$ is an argument $\langle \Psi, \psi \rangle$ where $\psi = \neg(\phi_1 \wedge \ldots \wedge \phi_n)$ and $\{\phi_1, \ldots, \phi_n\} \subseteq \Phi$. A *rebuttal* for an argument $\langle \Phi, \phi \rangle$ is an argument $\langle \Psi, \psi \rangle$ iff $\psi \Leftrightarrow \neg\phi$ is a tautology.

Argument A attacks argument B where $A = \langle \Psi, \psi \rangle$ and $B = \langle \Phi, \phi \rangle$ if either A undercuts B or A rebuts B.

Throughout the section, we will make the assumption that Ar is finite. Indeed, if Ar is the set of arguments that has been "received" by an agent, it is very reasonable to assume that the agent, who started

[1] Like in [37], we adopt the restriction to the finite case in order to use standard definitions of possibilistic logic. Extensions of possibilistic logic to the infinite case are discussed for example in [122].

operating at some time in the past and has a finite history, may have received, during its finite life, a finite number of arguments from finitely many sources.

Another assumption that we make is that an agent never forgets an argument it has been offered. Therefore, Ar may never shrink in time, i.e., if we denote by Ar_t the set of arguments received by an agent up to time t,

$$t_1 < t_2 \Rightarrow Ar_{t_1} \subseteq Ar_{t_2}. \tag{3.1}$$

Given an argument $A \in Ar$, we will denote by $\mathrm{src}(A)$ the set of the sources of A.

Fuzzy Labeling. In order to provide the intuition behind the idea of a fuzzy labeling of the arguments, consider the following dialogue in the context of a murder.

Example 28. The judge has to decide whether John has killed Mary. The agents are Wit_1 and Wit_2, two witnesses, a coroner *Cor* and the judge *Jud*. Assume the judge completely trusts the two witnesses but he does not quite trust (lower degree of trust) the coroner because it is well-known that he is almost always drunk. The judge starts with argument A: "If John did not kill Mary, then John is innocent" where the premise is "If John did not kill Mary" and the conclusion is "John is innocent". Then, the judge listens to the depositions of the two witnesses. Wit_1 asserts argument B: "I saw John killing Mary, thus John killed Mary". Argument B attacks A's premise so we have an attack $B \rightarrow A$. Wit_2 claims C: "John was at the theater with me when Mary was killed, thus John did not kill Mary". Argument C attacks B's conclusion and this leads to $C \rightarrow B$. Finally, the judge listens to the deposition of the coroner who asserts D: "Mary was killed before 6 p.m., thus when Mary was killed the show was still to begin". Argument D attacks C's premise introducing an attack $D \rightarrow C$. The attack relation is as follows: $D \rightarrow C, C \rightarrow B, B \rightarrow A$.

Example 28 presents a scenario where the arguments cannot be evaluated from the beginning in the same way because of the degree of trust assigned to their source. In order to account for the fact that arguments may originate from sources that are trusted only to a certain degree, we extend the (crisp) abstract argumentation structure described in Section 2.3 by allowing gradual membership of arguments in the set of arguments Ar. In other words, Ar is a fuzzy set of trustful arguments, and $Ar(A)$, the membership degree of argument A in Ar, is given by the trust degree of the most reliable (i.e., trusted) source that offers argument A[2],

$$Ar(A) = \max_{s \in \mathrm{src}(A)} \tau_s, \tag{3.2}$$

where τ_s is the degree to which source $s \in \mathrm{src}(A)$ is trusted.

It must be stressed that the fuzzy contribution in our approach is different from the one proposed by Janssen [179]. Their fuzzy approach enriches the expressive power of classical argumentation by allowing to represent the relative strength of the attack relations between the arguments, while in our approach the attack relations remains crisp; fuzzyness is introduced to represent uncertainty due to the fact that information sources can also be "just" partially trusted.

Parsons *et al.* [296] introduce a framework for decision making where they define trust-extended argumentation graphs in which each premise, inference rule and conclusion is associated to the trustworthiness degree of the source proposing it. Thus, given two arguments rebutting each others, the argument whose conclusion has an higher trust value is accepted. The difference is that in such a framework the "labels",

[2]Here, we suppose that the agent is optmistic. To represent the behaviour of a pessimistic agent, we should use the min operator, for example.

i.e., the trust values, associated to the arguments never change and the arguments are always accepted with the same degree even if they are attacked by more trusted arguments.

This fuzzification of Ar provides a natural way of associating strengths to arguments, and suggests rethinking the labeling of an argumentation framework in terms of fuzzy degrees of argument acceptability. Matt and Toni [217] define the strength of the argument the proponent embraces as his long run expected payoff. The difference with our fuzzy labeling is that they compute these strengths from probability distributions on the values of a game. The idea in common with our work is to replace the three-valued labeling with a graded labeling function.

Definition 55. (Fuzzy AF-labeling) Let $\langle Ar, \rightarrow \rangle$ be an abstract argumentation framework. A fuzzy AF-labeling is a total function $\alpha : Ar \rightarrow [0,1]$.

Such an α may also be regarded as (the membership function of) the fuzzy set of acceptable arguments: $\alpha(A) = 0$ means the argument is outright unacceptable, $\alpha(A) = 1$ means the argument is fully acceptable, and all cases inbetween are provided for.

Intuitively, the acceptability of an argument should not be greater than the degree to which the arguments attacking it are unacceptable:

$$\alpha(A) \leq 1 - \max_{B:B \rightarrow A} \alpha(B). \tag{3.3}$$

This is, indeed, a fuzzy reformulation of two basic postulates for reinstatement proposed by Caminada [84] to characterize the labeling of arguments: (1) an argument must be *in* iff all of its attackers are *out*; (2) an argument must be *out* iff there exists an *in* argument that attacks it.

Furthermore, it seems reasonable to require that

$$\alpha(A) \leq Ar(A), \tag{3.4}$$

i.e., an argument cannot be more acceptable than the degree to which its sources are trusted.

By combining the above two postulates, we obtain the following definition.

Definition 56. (Fuzzy Reinstatement Labeling) Let α be a fuzzy AF-labeling. We say that α is a fuzzy reinstatement labeling iff, for all arguments A,

$$\alpha(A) = \min\{Ar(A), 1 - \max_{B:B \rightarrow A} \alpha(B)\}. \tag{3.5}$$

We can verify that the fuzzy reinstatement labeling is a generalization of the crisp reinstatement labeling defined by Caminada, whose *in* and *out* labels are particular cases corresponding, respectively, to $\alpha(A) = 1$ and $\alpha(A) = 0$. What about the *undec* label in the fuzzy case? One might argue that it corresponds to $\alpha(A) = 0.5$; however, an exam of the case of two arguments attacking each other, $A \rightarrow B$ and $B \rightarrow A$, with $Ar(A) = Ar(B) = 1$, reveals that any fuzzy reinstatement labeling α must satisfy the equation

$$\alpha(A) = 1 - \alpha(B), \tag{3.6}$$

which has infinitely many solutions with $\alpha(A) \in [0,1]$. We can conclude that there are infinitely many degrees of "undecidedness" due to the trustworthiness of the source, of which 0.5 is but the most undecided representative. These degrees of "undecidedness" express how much the agent tends to accept those arguments proposed by not fully trusted agents.

Given a fuzzy argumentation framework, how to compute its fuzzy reinstatement labeling? The answer to this question amounts to solving a system of n non-linear equations, where $n = \|\text{supp}(Ar)\|$, i.e., the

number of arguments belonging to some non-zero degree in the fuzzy argumentation framework, of the same form as Equation 3.5, in n unknown variables, namely, the labels $\alpha(A)$ for all $A \in \text{supp}(Ar)$. Since iterative methods are usually the only choice for solving systems of non-linear equations, we will resort to this technique, but with an eye to how the labeling is computed in the crisp case. In particular, we draw some inspiration from Caminada [83]'s idea. We start with an all-in labeling (a labeling in which every argument is labeled with the degree it belongs to Ar). We introduce the notion of illegal labeling for argument A with respect to Definition 56.

Definition 57. (Illegal labeling) Let α be a fuzzy labeling and A be an argument. We say that A is illegally labeled iff $\alpha(A) \neq \min\{Ar(A), 1 - \max_{B:B\to A} \alpha(B)\}$.

In order to have an admissible labeling, the absence of illegally labeled arguments is required. As Caminada [83], we need a way of changing the illegal label of an argument, without creating other illegally labeled arguments.

We denote by $\alpha_0 = Ar$ the initial labeling, and by α_t the labeling obtained after the t^{th} iteration of the labeling algorithm.

Definition 58. Let α_t be a fuzzy labeling. An iteration in α_t is carried out by computing a new labeling α_{t+1} for all arguments A as follows:

$$\alpha_{t+1}(A) = \frac{1}{2}\alpha_t(A) + \frac{1}{2}\min\{Ar(A), 1 - \max_{B:B\to A}\alpha_t(B)\}. \tag{3.7}$$

Note that Equation 3.7 guarantees that $\alpha_t(A) \leq Ar(A)$ for all arguments A and for each step of the algorithm.

The above definition actually defines a sequence $\{\alpha_t\}_{t=0,1,\ldots}$ of labelings.

Theorem 16. *The sequence $\{\alpha_t\}_{t=0,1,\ldots}$ defined above converges.*

Proof. We have to prove that, for all A, there exists a real number $L_A \in [0, Ar(A)]$ such that, for all $\varepsilon > 0$, there exists N_A such that, for every $t > N_A$, $|\alpha_t(A) - L_A| < \varepsilon$.

The proof is quite straightforward if one assumes the attack relation to be acyclic. In that case, the thesis can be proved by structural induction on the attack relation: the basis is that if argument A is not attacked by any other argument, Equation 3.7 reduces to

$$\alpha_{t+1}(A) = \frac{1}{2}\alpha_t(A) + \frac{1}{2}Ar(A),$$

and, since $\alpha_0 = Ar(A)$, the sequence is constant and thus trivially converges to $Ar(A)$. The inductive step consists of assuming that $\{\alpha_t(B)\}_{t=0,1,\ldots}$ converges for all arguments B such that $B \to A$, and proving that then $\{\alpha_t(A)\}_{t=0,1,\ldots}$ converges as well. If all $\{\alpha_t(B)\}_{t=0,1,\ldots}$ converge, then so does $\{\mu_t(A)\}_t = \{\min\{Ar(A), 1 - \max_{B:B\to A}\alpha_t(B)\}\}_t$, i.e., there exists a real number $L_A \in [0, Ar(A)]$ such that, for all $\varepsilon > 0$, there exists N_A such that, for every $t > N_A$, $|\mu_t(A) - L_A| < \varepsilon$, or $L_A - \varepsilon < \mu_t(A) < L_A + \varepsilon$. Equation 3.7 reduces to

$$\alpha_{t+1}(A) = \frac{1}{2}\alpha_t(A) + \frac{1}{2}\mu_t(A),$$

We have to distinguish two cases. If $\alpha_{t+1}(A) \geq L_A$,

$$\begin{aligned}
|\alpha_{t+1}(A) - L_A| &= \alpha_{t+1}(A) - L_A = \\
&= \frac{1}{2}\alpha_t(A) + \frac{1}{2}\mu_t(A) - L_A < \\
&< \frac{1}{2}\alpha_t(A) + \frac{1}{2}(L_A + \varepsilon) - L_A = \\
&= \frac{1}{2}\alpha_t(A) - \frac{1}{2}L_A + \varepsilon/2 = \\
&= \frac{1}{2}(\alpha_t(A) - L_A) + \varepsilon/2 \leq \\
&\leq \frac{1}{2}|\alpha_t(A) - L_A| + \varepsilon/2.
\end{aligned}$$

Otherwise, $\alpha_{t+1}(A) < L_A$, and

$$\begin{aligned}
|\alpha_{t+1}(A) - L_A| &= L_A - \alpha_{t+1}(A) = \\
&= L_A - \frac{1}{2}\alpha_t(A) - \frac{1}{2}\mu_t(A) < \\
&< L_A - \frac{1}{2}\alpha_t(A) - \frac{1}{2}(L_A - \varepsilon) = \\
&= \frac{1}{2}L_A - \frac{1}{2}\alpha_t(A) + \varepsilon/2 = \\
&= \frac{1}{2}(L_A - \alpha_t(A)) + \varepsilon/2 \leq \\
&\leq \frac{1}{2}|\alpha_t(A) - L_A| + \varepsilon/2.
\end{aligned}$$

Therefore, $|\alpha_t(A) - L_A| < |\alpha_0(A) - L_A|2^{-t} + \varepsilon 2^{-t} \leq 2^{-t} + \varepsilon 2^{-t} = \varepsilon_1 + \varepsilon_2$.

The proof in the general case where attack cycles may exist is based on the idea that convergence in cycles may be proved separately, by assuming that $\{\alpha_t(B)\}_{t=0,1,\ldots}$ converges for all arguments B attacking any of the arguments in the cycle.

Let arguments $A_0, A_1, \ldots, A_{n-1}$ form a cycle, i.e., for all $i = 0, \ldots, n-1$, $A_i \to A_{i+1 \mod n}$, and let

$$u(A_i) = \min\{Ar(A_i), 1 - \max_{\substack{B:B\to A_i \\ B\notin\{A_0,\ldots,A_{n-1}\}}} L_B\}$$

be the upper bound of the feasible values for $\alpha(A_i)$. Note that a cycle with no external arguments attacking arguments of the cycle is a special case, whereby $u(A_i) = Ar(A_i)$ for all arguments in the cycle.

For every pair of arguments $(A_i, A_{i+1 \mod n})$, for α to be a fuzzy reinstatement labeling it should be

$$\alpha(A_{i+1 \mod n}) = \min\{u(A_{i+1 \mod n}), 1 - \alpha(A_i)\}$$

and

$$\sum_{i=0}^{n-1} \alpha(A_i) \leq \min\left\{\frac{n}{2}, \sum_{i=0}^{n-1} u(A_i)\right\}.$$

Now, if α_t is not yet a solution of Equation 3.5, there are two cases:

t	$\alpha_t(A)$	$\alpha_t(B)$	$\alpha_t(C)$
0	1	0.4	0.2
1	0.9	0.2	0.2
2	0.85	0.15	0.2
3	0.825	0.15	0.2
4	0.8125	0.1625	**0.2**
5	**0.8**	0.175	$\downarrow$
6	$\downarrow$	**0.2**	

Figure 3.12: Fuzzy labeling on AF: $A \to B, B \to C, C \to A$.

1. either $\sum_{i=0}^{n-1} \alpha_t(A_i) > \frac{n}{2}$, and there exists at least an argument A_i such that $\alpha_t(A_i) > 1 - \alpha_t(A_{i+1 \mod n})$; in this case, then,

$$\begin{aligned} \min\{u(A_i), 1 - \alpha_t(A_{i+1 \mod n})\} &\leq \\ 1 - \alpha_t(A_{i+1 \mod n}) &< \alpha_t(A_i) \end{aligned}$$

   and $\alpha_{t+1}(A_i) < \alpha_t(A_i)$, whence

$$\sum_{i=0}^{n-1} \alpha_{t+1}(A_i) < \sum_{i=0}^{n-1} \alpha_t(A_i);$$

2. or $\sum_{i=0}^{n-1} \alpha_t(A_i) < \frac{n}{2}$, and there exists at least an argument A_i such that

$$\alpha_t(A_i) < \min\{u(A_i), 1 - \alpha_t(A_{i+1 \mod n})\};$$

   but then $\alpha_{t+1}(A_i) > \alpha_t(A_i)$, whence

$$\sum_{i=0}^{n-1} \alpha_{t+1}(A_i) > \sum_{i=0}^{n-1} \alpha_t(A_i).$$

Therefore, α_t converges for all the arguments in the cycle, and this concludes the proof. □

An example of the calculation of the fuzzy labeling for an odd cycle with three arguments A, B, and C, such that $A \to B, B \to C, C \to A$ and $Ar(A) = 1$, $Ar(B) = 0.4$, and $Ar(C) = 0.2$, is presented in Figure 3.12.

We may now define the fuzzy labeling of a fuzzy argumentation framework as the limit of $\{\alpha_t\}_{t=0,1,\ldots}$.

Definition 59. Let $\langle Ar, \to \rangle$ be a fuzzy argumentation framework. A fuzzy reinstatement labeling for such argumentation framework is, for all arguments A,

$$\alpha(A) = \lim_{t \to \infty} \alpha_t(A). \tag{3.8}$$

The convergence speed of the labeling algorithm is linear, as the proof of convergence suggests: in practice, a small number of iterations is enough to get so close to the limit that the error is less than the precision with which the membership degrees are represented in the computer.

Example 29 (Continued)**.** Consider again the dialogue in the context of a murder. The judge fully trusts the two witnesses but he assigns a lower degree of trustworthiness to the coroner. The labels of the arguments at the beginning are: $\alpha(A) = Ar(A) = 1$, $\alpha(B) = Ar(B) = 1$, $\alpha(C) = Ar(C) = 1$, $\alpha(D) = Ar(D) = 0.3$. The fuzzy reinstatement labeling returns the following values: $\alpha(D) = 0.3$, $\alpha(C) = 0.7$, $\alpha(B) = 0.3$, and $\alpha(A) = 0.7$.

Belief revision. In the proposed framework, belief reinstatement is then guaranteed thanks to the integration of the argumentation framework with the belief-change phase. More precisely, when a new argument arrives, the argumentation framework is updated using the fuzzy labeling algorithm. Therefore, each argument reinstated by the algorithm will induce the reinstatement, to some extent, of the conclusion of the argument in the belief set and of all the formulas that logically follow from the belief set.

The membership function of a fuzzy set describes the more or less possible and mutually exclusive values of one (or more) variable(s). Such a function can then be seen as a possibility distribution [329]. If π_x is the fuzzy set of possible values of variable x, π_x is called the possibility distribution associated to x; $\pi_x(v)$ is the possibility degree of x being equal to v. A possibility distribution for which there exists a completely possible value ($\exists v_0 : \pi(v_0) = 1$) is said to be *normalized*.

Definition 60. (Possibility and Necessity Measures) A possibility distribution π induces a *possibility measure* and its dual *necessity measure*, denoted by Π and N respectively. Both measures apply to a crisp set A and are defined as follows:

$$\Pi(A) = \sup_{s \in A} \pi(s); \tag{3.9}$$

$$N(A) = 1 - \Pi(\bar{A}) = \inf_{s \in \bar{A}} \{1 - \pi(s)\}. \tag{3.10}$$

As convincingly argued by [117], a *belief* should be regarded as a necessity degree induced by a normalized possibility distribution

$$\pi : \Omega \to [0, 1], \tag{3.11}$$

which represents a plausibility order of possible states of affairs: $\pi(\mathscr{I})$ is the possibility degree of interpretation $\mathscr{I}$.

Starting from such an insight, a fuzzy reinstatement labeling α determines a set of beliefs in a natural way. Given argument $A = \langle \Phi, \phi \rangle$, let $\mathrm{con}(A)$ denote the conclusion of A, i.e., $\mathrm{con}(\langle \Phi, \phi \rangle) = \phi$. The possibility distribution π induced by a fuzzy argumentation framework may be constructed by letting, for all interpretation $\mathscr{I}$,

$$\pi(\mathscr{I}) = \min\{1, 1 + \max_{A:\mathscr{I} \models \mathrm{con}(A)} \alpha(A) - \max_{B:\mathscr{I} \not\models \mathrm{con}(B)} \alpha(B)\}. \tag{3.12}$$

The first maximum in the above equation accounts for the most convincing argument compatible with world $\mathscr{I}$, whereas the second maximum accounts for the most convincing argument against world $\mathscr{I}$. A world will be possible to an extent proportional to the difference between the acceptability of the most convincing argument supporting it and the acceptability of the most convincing argument against it. The world will be considered completely possible if such difference is positive or null, but it will be considered less and less possible (or plausible) as such difference grows more and more negative.

Theorem 17. *Any π defined as per Equation 3.12 is normalized.*

Proof. Either $\pi(\mathscr{I}) = 1$ for all $\mathscr{I}$, and π is trivially normalized, or there exists an interpretation, say $\mathscr{I}_0$, such that $\pi(\mathscr{I}_0) < 1$. By Equation 3.12, then, it must be

$$\max_{A:\mathscr{I}_0 \models \mathrm{con}(A)} \alpha(A) < \max_{B:\mathscr{I}_0 \not\models \mathrm{con}(B)} \alpha(B).$$

But then, let us consider the complementary interpretation $\overline{\mathscr{I}_0}$, which maps all atoms to a truth value that is the opposite of the truth value they are mapped to by $\mathscr{I}_0$. Clearly, all formulas satisfied by $\mathscr{I}_0$ are not satisfied by $\overline{\mathscr{I}_0}$ and *vice versa*. Therefore,

$$\pi(\overline{\mathscr{I}_0}) = \min\{1, 1 + \max_{B:\mathscr{I} \not\models \mathrm{con}(B)} \alpha(B) - \max_{A:\mathscr{I} \models \mathrm{con}(A)} \alpha(A)\} = 1.$$

In other words, if a world is not completely plausible, its opposite must be completely plausible, and for this reason π is always normalized. □

The degree to which a given arbitrary formula $\phi \in \mathscr{L}$ is believed can be calculated from the possibility distribution induced by the fuzzy argumentation framework as

$$\mathbf{B}(\phi) = N([\phi]) = 1 - \max_{\mathscr{I} \not\models \phi} \{\pi(\mathscr{I})\}. \tag{3.13}$$

Such $\mathbf{B}$ may be regarded, at the same time, as a fuzzy modal epistemic operator or as a fuzzy subset of $\mathscr{L}$.

A powerful feature of such an approach based on a possibility distribution is that $\mathbf{B}(\phi)$ can be computed for any formula ϕ, not just for formulas that are the conclusion of some argument. For instance, if A is an argument whose conclusion is p and B is an argument whose conclusion is $p \supset q$, and $\alpha(A) = \alpha(B) = 1$, then not only $\mathbf{B}(p) = \mathbf{B}(p \supset q) = 1$, but also $\mathbf{B}(q) = 1$, $\mathbf{B}(p \wedge q) = 1$, etc.

Straightforward consequences of the properties of possibility and necessity measures are that $\mathbf{B}(\phi) > 0 \Rightarrow \mathbf{B}(\neg\phi) = 0$, this means that if the agent somehow believes ϕ then it cannot believe $\neg\phi$ at all;

$$\begin{aligned}
\mathbf{B}(\top) &= 1, && (3.14)\\
\mathbf{B}(\bot) &= 0, && (3.15)\\
\mathbf{B}(\phi \wedge \psi) &= \min\{\mathbf{B}(\phi), \mathbf{B}(\psi)\}, && (3.16)\\
\mathbf{B}(\phi \vee \psi) &\geq \max\{\mathbf{B}(\phi), \mathbf{B}(\psi)\}. && (3.17)
\end{aligned}$$

We can finally investigate the degree of the agent's belief in terms of the labeling values of the arguments. Let A, B, A_0, and B_0 represent arguments, and let $\mu \in (0,1]$ be a degree of belief. Then, for all $\phi \in \mathscr{L}$,

$$\begin{aligned}
& \mathbf{B}(\phi) \geq \mu \\
\Leftrightarrow\ & \forall \mathscr{I} \not\models \phi \quad \pi(\mathscr{I}) \leq 1 - \mu, \text{ (Eq. 3.13)} \\
\Leftrightarrow\ & \forall \mathscr{I} \not\models \phi \\
& 1 + \max_{A:\mathscr{I} \models \mathrm{con}(A)} \alpha(A) - \max_{B:\mathscr{I} \not\models \mathrm{con}(B)} \alpha(B) \leq 1 - \mu, \text{ (Eq. 3.12)} \\
\Leftrightarrow\ & \forall \mathscr{I} \not\models \phi \max_{B:\mathscr{I} \not\models \mathrm{con}(B)} \alpha(B) - \max_{A:\mathscr{I} \models \mathrm{con}(A)} \alpha(A) \geq \mu, \\
\Leftrightarrow\ & \forall \mathscr{I} \not\models \phi\, \exists B_0 : \mathscr{I} \not\models \mathrm{con}(B_0), \forall A : \mathscr{I} \models \mathrm{con}(A), \\
& \alpha(B_0) - \alpha(A) \geq \mu.
\end{aligned}$$

In words, a necessary and sufficient condition for formula ϕ to be believed to some extent is that, for all interpretation $\mathscr{I}$ which does not satisfy ϕ, there exists an argument whose consequence is not satisfied by $\mathscr{I}$ that is more accepted than every argument whose consequence is satisfied by $\mathscr{I}$.

Therefore, the necessary and sufficient condition for formula ϕ not to be believed may be stated as follows:

$$\begin{aligned} & \mathbf{B}(\phi) = 0 \\ \Leftrightarrow \quad & \exists \mathscr{I}_0 \not\models \phi, \exists A_0 : \mathscr{I}_0 \models \mathrm{con}(A_0), \forall B : \mathscr{I}_0 \not\models \mathrm{con}(B), \\ & \alpha(A_0) \geq \alpha(B). \end{aligned}$$

Indeed,

$$\begin{aligned} & \mathbf{B}(\phi) = 0 \\ \Leftrightarrow \quad & \exists \mathscr{I}_0 \not\models \phi : \pi(\mathscr{I}_0) = 1, \\ \Leftrightarrow \quad & \exists \mathscr{I}_0 \not\models \phi : \\ & \min\{1, 1 + \max_{A:\mathscr{I}_0 \models \mathrm{con}(A)} \alpha(A) - \max_{B:\mathscr{I}_0 \not\models \mathrm{con}(B)} \alpha(B)\} = 1, \\ \Leftrightarrow \quad & \exists \mathscr{I}_0 \not\models \phi : \max_{A:\mathscr{I}_0 \models \mathrm{con}(A)} \alpha(A) \geq \max_{B:\mathscr{I}_0 \not\models \mathrm{con}(B)} \alpha(B). \end{aligned}$$

In this case, a formula ϕ is not (or no more) believed by the agent iff there exists an interpretation $\mathscr{I}_0$ which does not satisfy ϕ and it is such that there exists an argument whose consequence is satisfied by $\mathscr{I}_0$ and is more accepted than all the arguments whose consequence is not satisfied by $\mathscr{I}_0$.

Therefore, if belief in ϕ is lost due to the arrival of an argument A which causes the labeling to change so that $\mathbf{B}(\phi) = 0$, a sufficient condition for reinstatement of ϕ is that another argument A' arrives causing the labeling to change so that $\mathbf{B}(\phi) > 0$. However, this does not mean that the previous labeling must be restored, but that it is enough that, for all $\mathscr{I} \not\models \phi$, there exists an argument $B_{\mathscr{I}}$ whose consequence is not satisfied by $\mathscr{I}$, such that $\alpha(B_{\mathscr{I}}) > \alpha(C)$, for all arguments C whose consequence is satisfied by $\mathscr{I}$.

Example 30 (Continued)**.** Suppose the judge finds that Wit_1 is little reliable since he was in love with Mary before they broke up because of John. The starting label of argument B becomes $\alpha(B) = Ar(B) = 0.2$. The degree to which $\mathrm{con}(A)$ is believed at the beginning of the dialogue is $\mathbf{B}(\mathrm{con}(A)) = 1$. Then the other three arguments are put forward and the fuzzy reinstatement labeling returns the following values after 53 iterations: $\alpha(D) = 0.3$, $\alpha(C) = 0.7$, $\alpha(B) = 0.2$, and $\alpha(A) = 0.8$. The condition for reinstatement of $\mathrm{con}(A)$ is that argument C causes the labeling to change such that $\mathbf{B}(\mathrm{con}(A)) > 0$. At the end, the judge believes in John's innocence with a degree given by $\mathbf{B}(\mathrm{con}(A)) = 0.8$.

Evaluation. We study now the behavior and the performances of the fuzzy-labeling algorithm over a benchmark for abstract argumentation, and then we report about the obtained results.

The aim of our experimental analysis is to assess the scalability of the fuzzy-labeling algorithm concerning two perspectives:

- the number of iterations needed for convergence with respect to the number of the nodes in the graph, and
- the time needed for convergence with respect to the number of the nodes in the graph.

It must be stressed that the time needed for convergence depends on (i) the time needed for computing each iteration, (ii) the time needed to update the α of each single argument, and (iii) the number of iterations required for the labeling to converge.

The benchmark we used to evaluate the performances of the fuzzy labeling algorithm is composed of different datasets for abstract argumentation tasks used in the literature. More precisely, we have considered the following datasets:

- The Perugia dataset [46, 45, 47]:[3] the dataset is composed of randomly generated directed-graphs. To generate random graphs, they adopted two different libraries. The first one is the Java Universal Network/Graph Framework (JUNG), a Java software library for the modeling, generation, analysis and visualization of graphs. The second library they used is NetworkX, a Python software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. Three kinds of networks are generated:
  - In the Erdős-Rényi graph model, the graph is constructed by randomly connecting n nodes. Each edge is included in the graph with probability p independent from every other edge. For the generation of these argumentation graphs, they adopted $p = c \log n/n$ (with c empirically set to 2.5), which ensures the connectedness of such graphs.
  - The Kleinberg graph model adds a number of directed long-range random links to an $n \times n$ lattice network, where vertices are the nodes of a grid with undirected edges between any two adjacent nodes. Links have a non-uniform distribution that favors edges to close nodes over more distant ones.
  - In the Barabási-Albert graph model, at each time step, a new vertex is created and connected to existing vertices according to the principle of "preferential attachment", such that vertices with higher degree have a higher probability of being selected for attachment.

  For more details about the generation of these networks as well as the graph models, we refer the reader to [46, 45, 47].

- The dataset used by Cerutti *et al.* in their KR 2014 paper [97] (which we will call the KR dataset): the dataset has been generated to evaluate a meta-algorithm for the computation of preferred labelings, based on the general recursive schema for argumentation semantics called SCC-Recursiveness. The dataset is composed of three sets of argumentation frameworks, namely:
  - 790 randomly generated argumentation frameworks where the number of strongly connected components (SCC) is 1, varying the number of arguments between 25 and 250 with a step of 25.
  - 720 randomly generated argumentation frameworks where the number of strongly connected components varies between 5 and 45 with a step of 5. The size of the SCCs is determined by normal distributions with means between 20 and 40 with a step of 5, and with a fixed standard deviation of 5. They similarly varied the probability of having attacks between arguments among SCCs.
  - 2800 randomly generated argumentation frameworks where the number of strongly connected components is between 50 and 80 with a step of 5.

[3] The dataset is available at `http://www.dmi.unipg.it/conarg/dwl/networks.tgz`.

$a \to b, b \to a$
$a \to b, b \to a, b \to c$
$a \to b, b \to a, b \to c, c \to d, d \to c$
$a \to b, b \to c, c \to d, d \to e, e \to f, f \to e$
$a \to b, b \to c, c \to d, d \to c$
$a \to b, b \to c, c \to a$
$a \to b, b \to a, b \to c, c \to c$
$a \to b, b \to a, b \to c, c \to c, d \to d$
$a \to b, b \to a, b \to c, a \to c, c \to d, d \to c$
$a \to b, b \to a, b \to c, a \to c, c \to d$
$a \to b, b \to c, c \to a, b \to d, a \to d, c \to d, d \to e, e \to d$
$a \to b, b \to c, e \to c, c \to d$
$a \to b, b \to a, b \to c, c \to d, d \to e, e \to c$
$a \to b, b \to c, c \to c$
$a \to b, b \to c, c \to a, a \to d, b \to d, c \to d$
$a \to b, a \to c, c \to a, c \to d, d \to c, d \to a, a \to d, c \to e, d \to f$

Figure 3.13: The "patterns" used for constructing the Sophia Antipolis dataset.

- The dataset presented by Vallati *et al.* at ECAI 2014 [303] (which we will call the ECAI dataset): the dataset was produced to study the features of argumentation frameworks. More precisely, it is composed of $10,000$ argumentation frameworks generated using a parametric random approach allowing to select (probabilistically - average, standard deviation) the density of attacks for each strongly connected component, and how many arguments (probabilistically) in each SCC attack how many arguments (probabilistically) in how many (probabilistically) other SCCs. The number of arguments ranges between 10 and $40,000$, and they exploited a 10-fold cross-validation approach on a uniform random permutation of the instances.

The availability of real-world benchmarks for argumentation problems is quite limited, with some few exceptions like [73] or AIFdb.[4] However, these benchmarks are tailored towards problems of argument mining and their representation as abstract argumentation frameworks usually leads to topologically simple graphs, such as cycle-free graphs. These kinds of graphs are not suitable for evaluating the computational performance of solvers for abstract argumentation problems. For this reason, we decided to use artificially generated graphs as benchmarks, in line with the preliminary performance evaluation of Bistarelli *et al.* [46].

In order to ensure the consideration of all kinds of interesting "patterns" that could appear in argumentation frameworks (e.g., the abstract argumentation frameworks used to exemplify the behaviour of the semantics in [19]), we have generated further graphs by composing these basic well-known examples of *interesting* argumentation patterns (shown in Figure 3.13) into bigger frameworks.

Our generated dataset (which we will call the Sophia Antipolis dataset)[5] consists of 20,000 argumentation graphs created through a random aggregation of the patterns shown above. This process has been executed with different settings in order to obtain complex graphs of specific sizes. In particular, a set of 1,000 argumentation graphs is generated for graph sizes from 5,000 to 100,000 nodes, with incremental steps of 5,000 nodes each. The aggregation of patterns has been done incrementally, and the connections (edges)

[4] `http://corpora.aifdb.org`

[5] The Sophia Antipolis dataset is available at `https://goo.gl/pN1M9r`.

between single patterns were generated randomly. The number of created graphs and their different sizes should support the evaluation of argumentation reasoning algorithms under a broad number of scenarios.

Figures 3.14–3.23 summarize the behavior of the fuzzy-labeling algorithm on the four datasets we considered. For each dataset, we applied the algorithm to the argumentation graphs with all argument weights set to 1 (i.e., arguments coming from fully trusted sources) and with random weights (i.e., arguments coming from a variety of more or less trusted sources as it may be the case in application scenarios like multiagent systems). From a first inspection of the figures, it is clear that certain graph types are harder than others: the Sophia Antipolis appears to be the hardest, followed by the Erdős-Rényi, the Barabási-Albert, and the KR + EKAI datasets. The Kleinberg dataset appears to be the easiest. Furthermore, for all datasets, the graphs with random weights never require a smaller number of iterations for convergence than their counterparts with all weights fixed to 1.

Figures 3.14 and 3.15 show the behaviour of the fuzzy labeling algorithm when applied to the Barabási-Albert dataset. In particular, Figure 3.14 (left-hand side) illustrates the evolution of the number of iterations needed to reach convergence when all the weights are equal to 1. We can notice that the curve follows a logarithmic rise with the increasing of the number of nodes. The figure illustrated through the (right-hand side) curve represents the evolution of the time needed to reach the convergence. It shows a behaviour rather linear. However, we can notice that the slope of the curve decreases with the increasing of the number of nodes. A similar behaviour is depicted in Figure 3.15 which illustrates the evolution of the quantity of time (in ms) needed to reach the convergence when the weights are assigned randomly. These two illustrations clearly show the capability of the fuzzy labeling algorithm to handle a growing amount of data.

In Figures 3.16 and 3.17, we present the behaviour of the fuzzy labeling algorithm when applied to the Erdős-Rényi dataset. We can notice that when all the weights are equal to 1, the convergence is reached very quickly both when considering the number of iterations, and the quantity of time needed for convergence. However, while such a quantity is quite similar with respect to the case in which the weights are randomly assigned, we can notice that the number of iterations needed for convergence is higher with respect to the behaviour illustrated in Figure 3.16. This can be due to the fact that the Erdős-Rényi dataset is constructed by randomly connecting the nodes. As we can see in Figures 3.18 and 3.19, the convergence with the Kleinberg dataset is even globally faster, either when all weights are equal to 1 or when the weights are randomly assigned. Instead, the behaviour on the Sophia Antipolis dataset, shown in Figures 3.22 and 3.23, is quite similar to the behaviors obtained with the Barabási-Albert dataset.

It is less evident, but the fuzzy-labeling algorithm behaves on the KR + ECAI dataset (illustrated in Figures 3.20 and 3.21) much like it does on the Barabási-Albert and Sophia Antipolis datasets, with the exception of a few small graphs which are outliers and which demand a relatively large number of iterations to converge. Nevertheless, the time behavior of Barabási-Albert, Sophia Antipolis and KR + ECAI is qualitatively identical.

Despite the differences among the various graph types, we have a rate of increase in time which is at most log-linear for all graph types and for all weight assignments.

3.4 Related work

Cognitive model of conflicts in trust using argumentation. Dix et al. [127] present trust as a major issue concerning the research challenges for argumentation. The question *Which agents are trustworthy?* is important for taking decisions and weighing the arguments of the other agents.

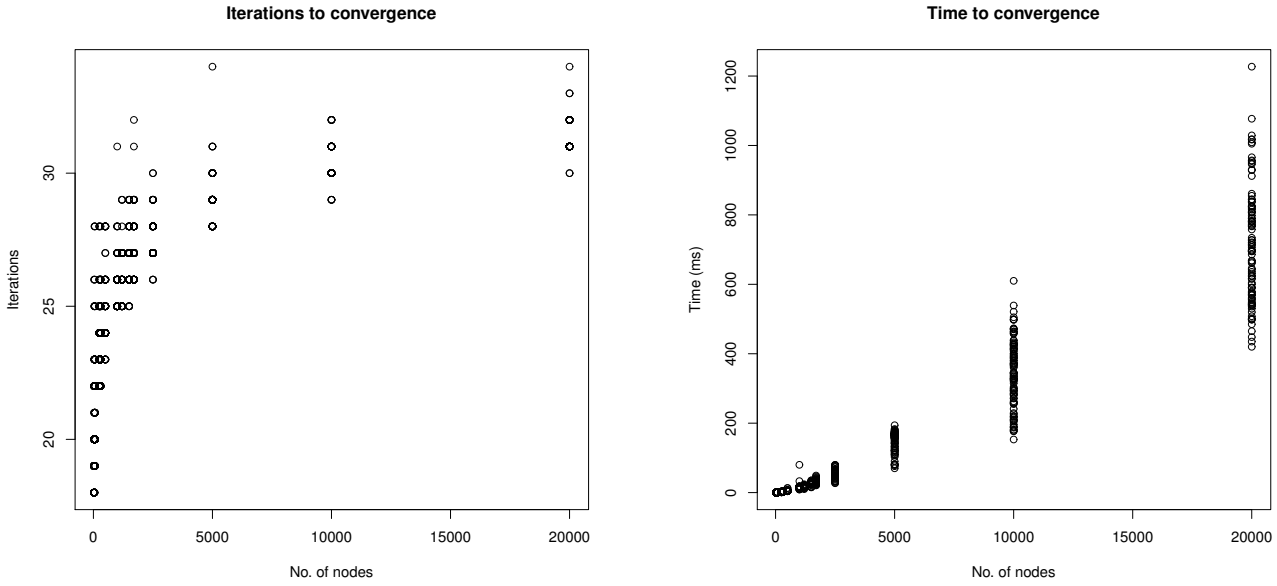


Figure 3.14: Barabási-Albert dataset of the Perugia benchmark with all weights equal to 1.0.

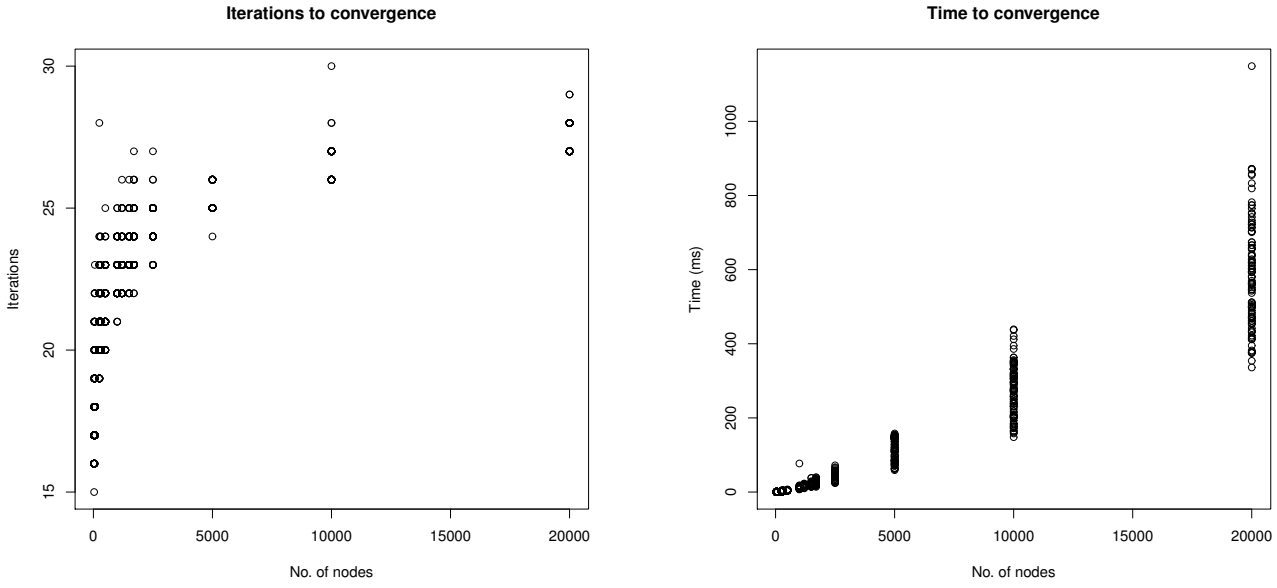


Figure 3.15: Barabási-Albert dataset of the Perugia benchmark with random weights.

Also Parsons *et al.* [252] present the provenance of trust as one of the mechanisms to be investigated in argumentation. They claim that a problem, particularly of abstract approaches such as Dung [133], is that they cannot express the provenance of trust, and the fact that argument b is attacked because b

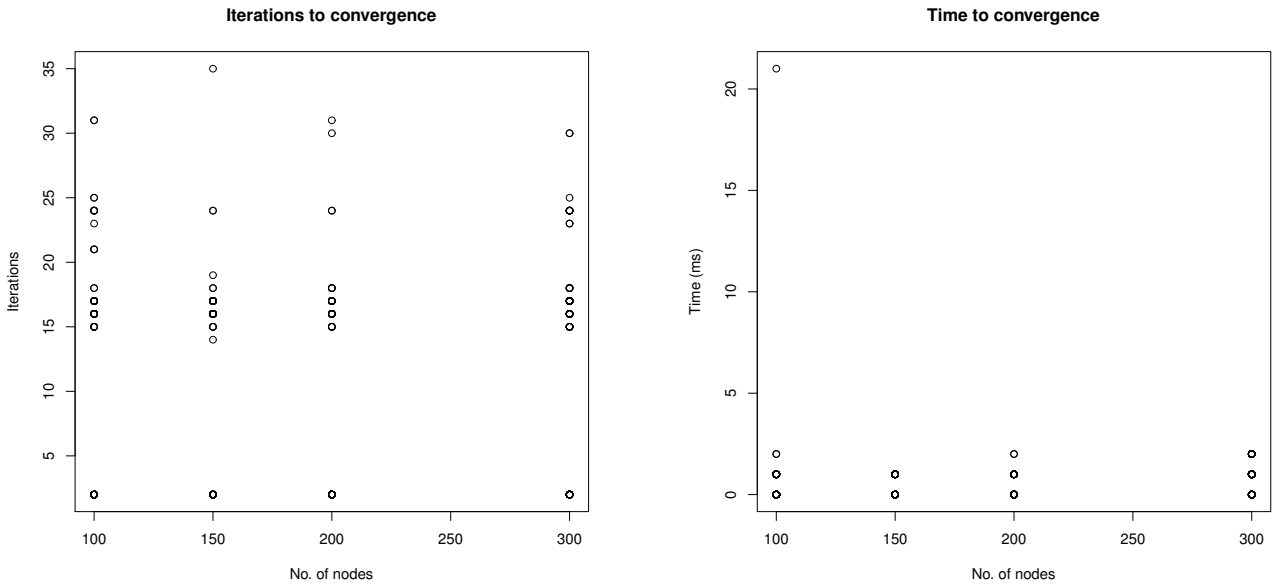


Figure 3.16: The Erdős-Rényi dataset of the Perugia benchmark with all weights equal to 1.0.

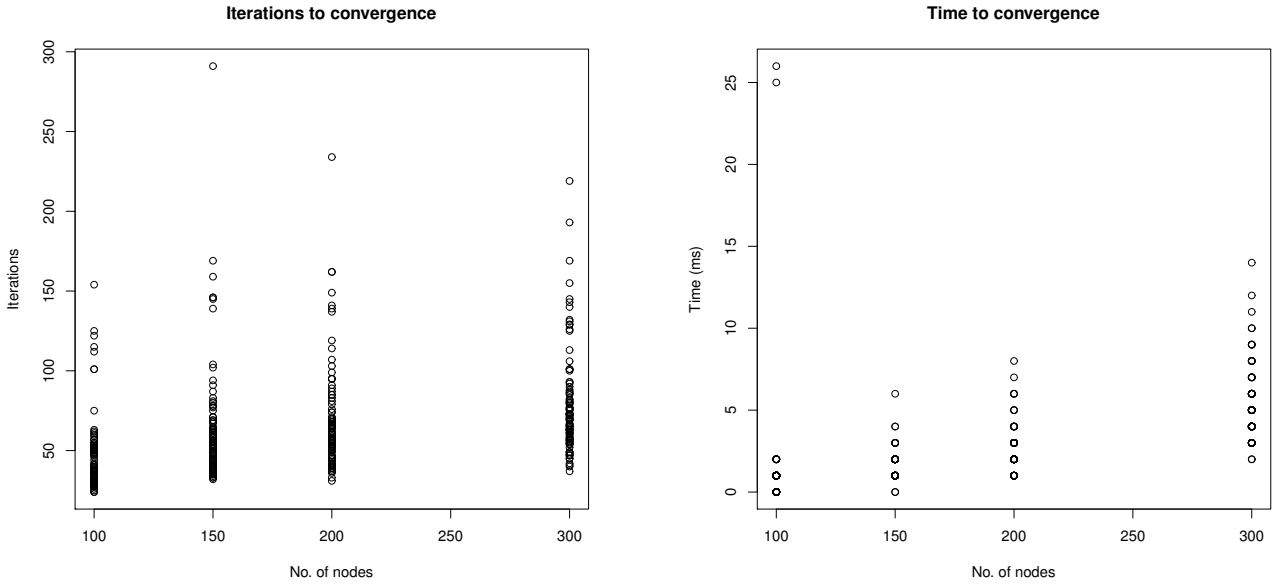


Figure 3.17: The Erdős-Rényi dataset of the Perugia benchmark with random weights.

is proposed by source s, who is not trustworthy. Starting from this observation, we propose a model of argumentation where the arguments are related to the sources and their acceptability is computed on the basis of the trustworthiness of the sources. Furthermore, our approach goes beyond this observation by

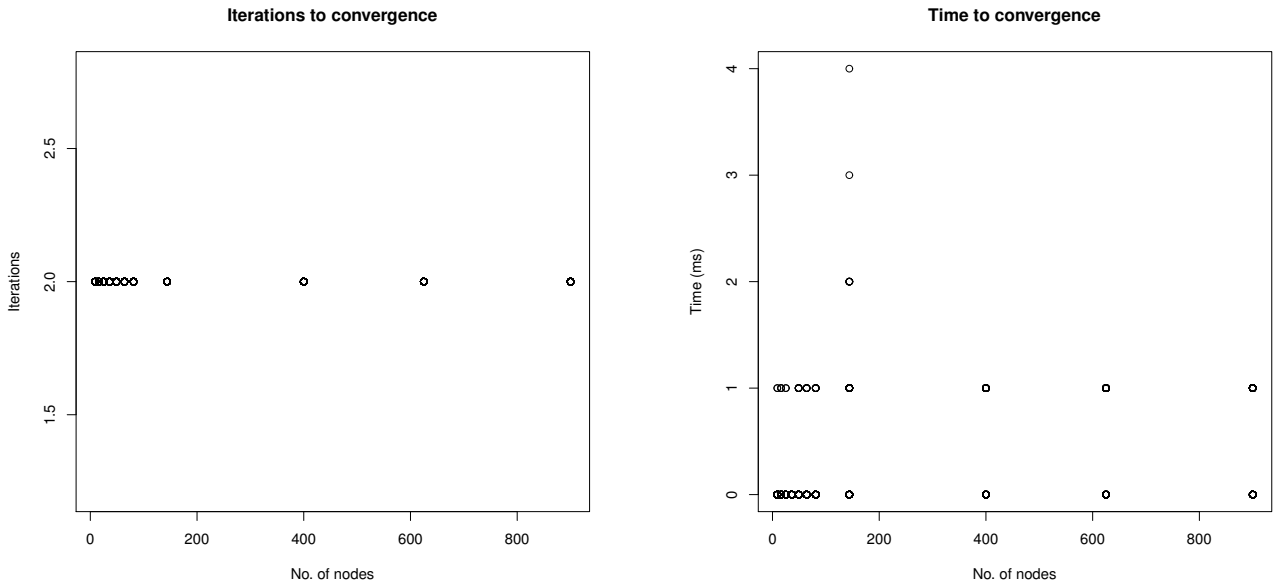


Figure 3.18: The Kleinberg dataset of the Perugia benchmark with all weights equal to 1.0.

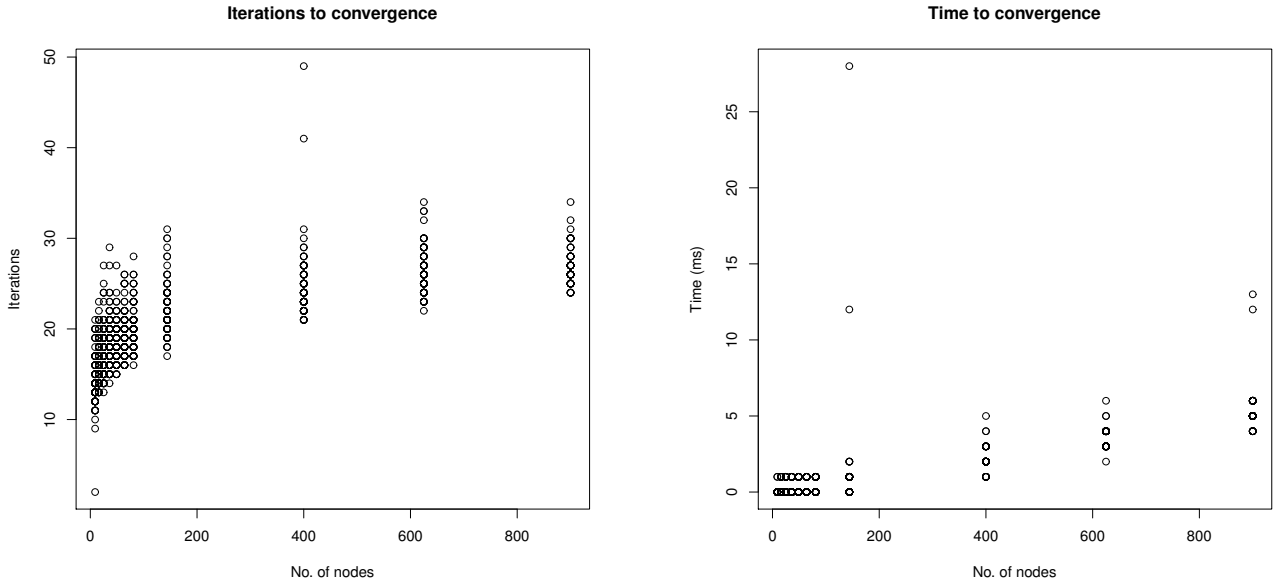


Figure 3.19: The Kleinberg dataset of the Perugia benchmark with random weights.

providing a feedback such that the final quality of the arguments influences the source evaluation as well.

Stranders et al. [293] propose an approach to trust based on argumentation that aims at exposing the rationale behind such trusting decisions. The aim of our work is different: we are interested in evaluating

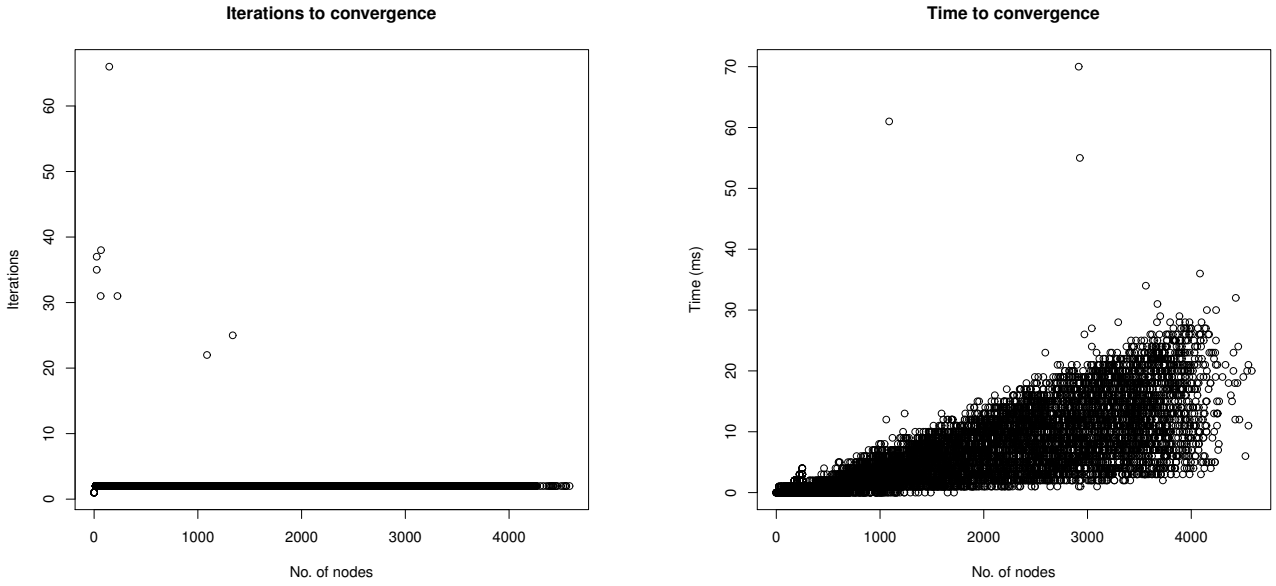


Figure 3.20: The benchmark consisting of the KR + ECAI dataset with all weights equal to 1.0.

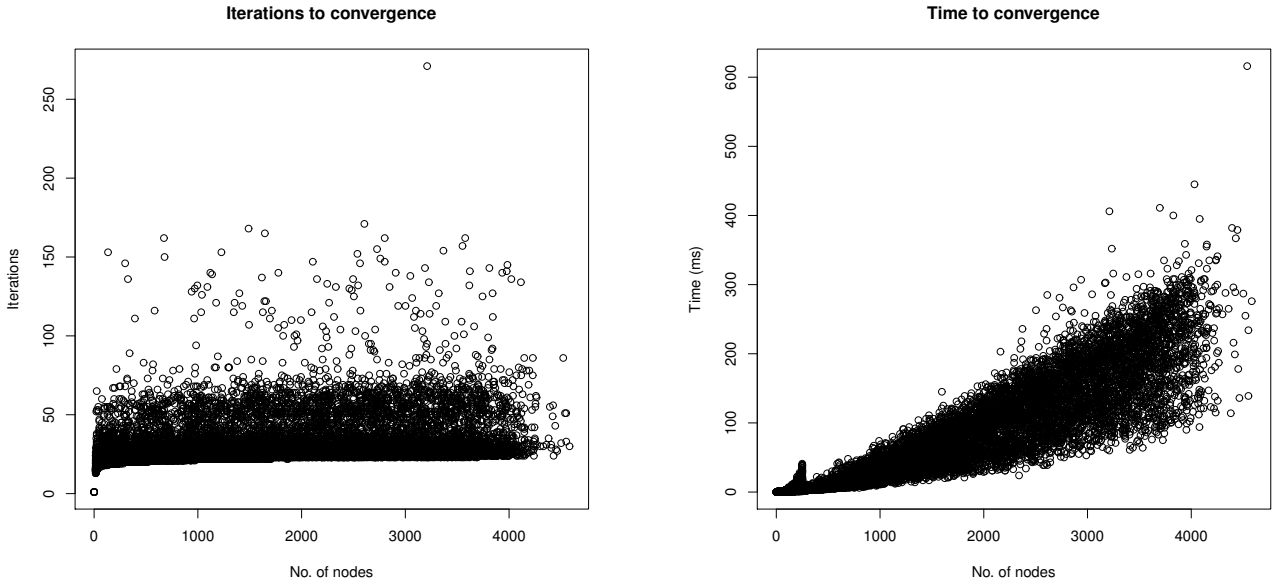


Figure 3.21: The benchmark consisting of the KR + ECAI dataset with random weights.

the arguments proposed by the sources with respect to their trustworthiness, instead of explaining, thanks to argumentation theory, the decisions about trusting or not another agent.

Prade [261] presents a bipolar qualitative argumentative modeling of trust where trust and distrust are

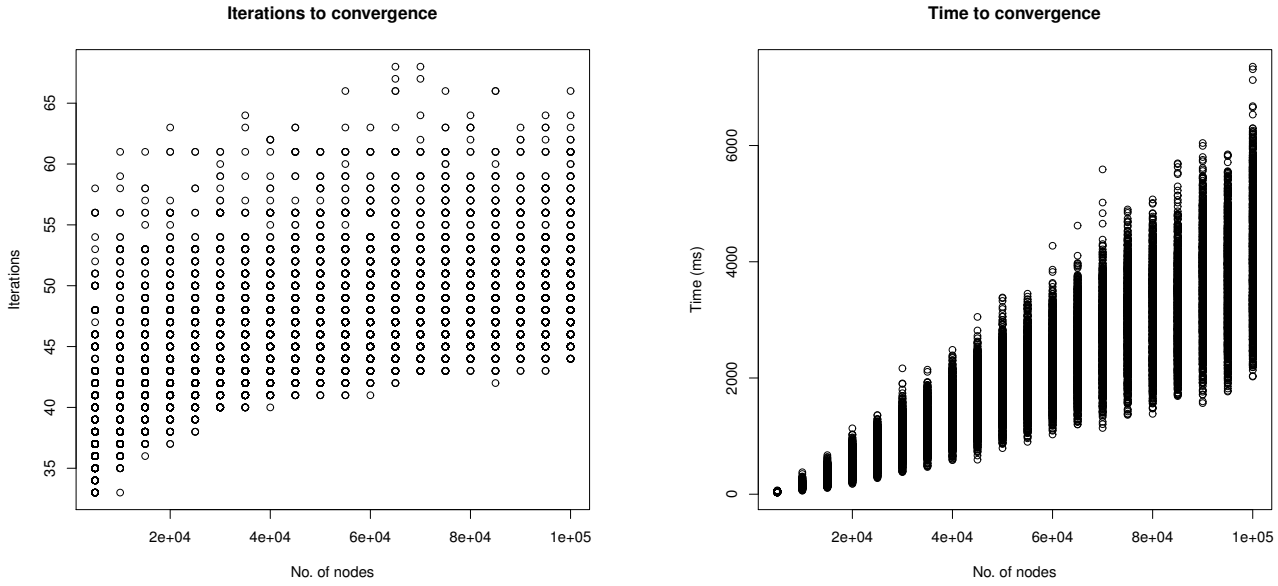


Figure 3.22: The Sophia Antipolis benchmark with all weights equal to 1.0.

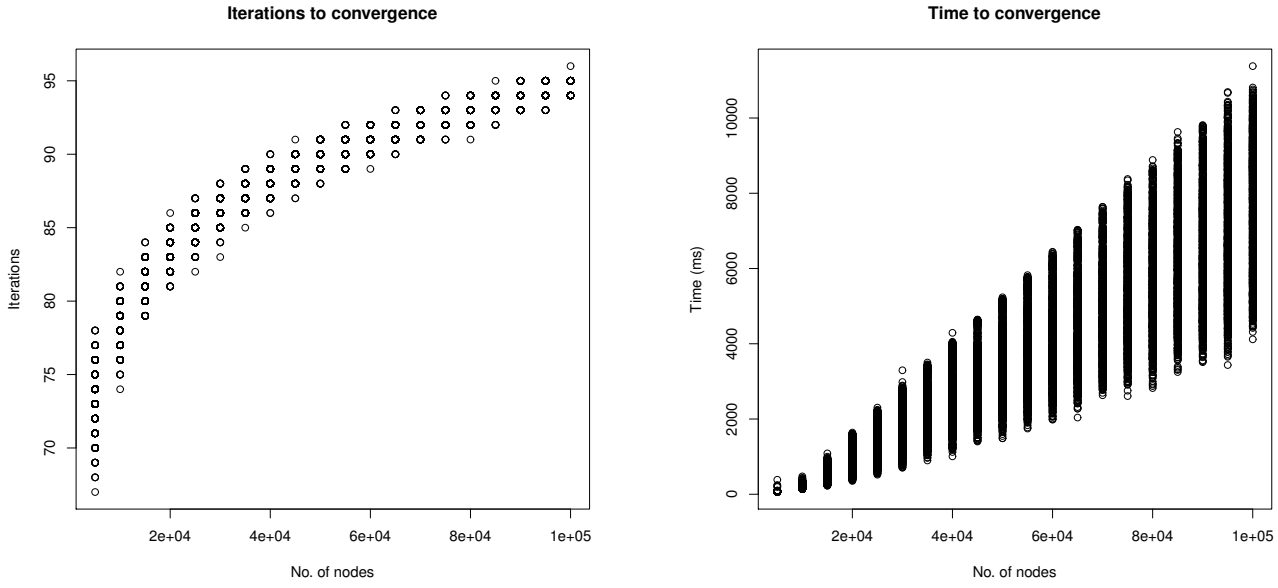


Figure 3.23: The Sophia Antipolis benchmark with random weights.

assessed independently. The author introduces also a notion of reputation which is viewed as an input information used by an agent for revising or updating his trust evaluation. Reputation contributes to provide direct arguments in favor or against a trust evaluation. In this chapter, we do not use observed behavior and

reputation to compute the starting trust value, and we model the socio-cognitive dynamics of trust such as the feedback and the trust dimensions, differently from [261].

Matt *et al.* [216] propose to construct a Dempster-Shafer belief function both from statistical data and from arguments in the context of contracts. We do not have arguments expressing the trustworthiness degree assigned to the other agents, but we accept the arguments depending on the trustworthiness of their sources. Moreover, in our model, the trustworthiness assigned to the arguments feeds back to the sources dynamically changing their own trustworthiness. We distinguish also the two dimensions of sincerity and competence.

Tang *et al.* [296] and Parsons *et al.* [253] present a framework to introduce the sources in argumentation and to express the degrees of trust. They define trust-extended argumentation graphs in which each premise, inference rule, and conclusion is associated to the trustworthiness degree of the source proposing it. Thus, given two arguments rebutting each other, the argument whose conclusion has a higher trust value is accepted. They do not have the possibility to directly attack the trustworthiness of the sources as well as the trustworthiness of single arguments and attacks. Again, the feedback towards the source as well as the distinction between competence and sincerity is not considered. We do not express the degrees of trust in a fine-grained way as done in [296, 253].

A huge amount of research has been conducted on trust, and some of these works are described below, even if in this chapter we limit our attention to the cognitive trust model of Castelfranchi and Falcone [90].

Castelfranchi and Falcone [90] stress the importance of this explicit cognitive account for trust in three ways. First, they criticize the game-theoretic view of trust which is prisoner of the Prisoner Dilemma mental frame, and reduce trust simply to a probability or perceived risk in decisions. Second, they find the quantitative aspects of trust (its strength or degree) on those mental ingredients (beliefs and goals) and on their strength. Third, they claim that this cognitive analysis of trust is fundamental for distinguishing among very different strategies for building or increasing trust; for founding mechanisms of image, reputation, persuasion, and argumentation in trust building. Apart from the cognitive model of Castelfranchi and Falcone [90] that define trust as "*a mental state, a complex attitude of an agent x towards another agent y about the behaviour/action a relevant for the goal g*", many other definitions have been provided in the literature.

In sociology, Gambetta [153] states that "*trust is the subjective probability by which an individual A expects that another individual B performs a given action on which its welfare depends*". Castelfranchi and Falcone [90] observe that this definition is correct. However, it is also quite a poor definition, since it just refers to one dimension of trust, i.e., predictability, while ignoring the "competence" dimension; it does not account for the meaning of "I trust B" where there is also the decision and the act of relying on B; and it does not explain what is such an evaluation made of and based on. Common elements of these definitions are a consistent degree of uncertainty and conflicting information associated with trust.

Another approach to model trust using modal logic is proposed by Lorini and Demolombe [212] where they present a concept of trust that integrates the truster's goal, the trustee's action ensuring the achievement of the truster's goal, and the trustee's ability and intention to do this action. In this chapter, we do not refer to the actions of the sources, but we provide a model for representing the conflicts the sources have to deal with trust. The introduction of the actions in our cognitive model is left as future work, and it will allow also to model willingness (source s should think that source p not only is able and can do that action/task, but p actually will do what s needs). In this chapter, we model only the competence and sincerity mental states of trust.

Another proposal is presented by Liau [202], in which the influence of trust on the assimilation of information into the source's mind is considered. The idea is that "if agent i believes that agent j has told him the truth on p, and he trusts the judgement of j on p, then he will also believe p". Extending the model by introducing goals to model the presented definitions is left for future work.

Wang and Singh [321], instead, understand trust in terms of belief and certainty: A's trust in B is reflected in the strength of A's belief that B is trustworthy. They formulate certainty in terms of evidence based on a statistical measure defined over a probability distribution of positive outcomes. Both Liau [202] and Wang and Singh [321] capture intuitions that play a role also in our approach, but they propose a simplified model of the nature and dynamics of trust, as opposed to the socio-cognitive model discussed by Castelfranchi [90].

Fuzzy labeling algorithm. Despite the existence of such a clear complementarity between these two fields of Artificial Intelligence, there are few works integrating them in a unitary multiagent framework. However, a consensus exists on the opportunity of integrating belief revision and argumentation. Cayrol *et al.* [95] do not integrate belief revision and argumentation, but propose a work on "revision in argumentation frameworks" in which they transpose the basic issue of revision into argumentation theory. They study the impact of the arrival of a new argument on the set of extensions of an abstract argumentation framework. Quignard *et al.* [264] describe a model for argumentation in agent interaction that shows how belief conflicts may be resolved by considering the relations between the agents' cognitive states and their choice of relevant argumentation moves. Paglieri *et al.* [246] claim that belief revision and argumentation can be seen as, respectively, the cognitive and social sides of the same epistemic coin and propose a preliminary framework which follows Toulmin's layout of argumentation. Falappa and colleagues [143] survey relevant work combining belief revision and argumentation. Besides, they develop a conceptual view on argumentation and belief revision as complementary disciplines used to explain reasoning. They propose four basic steps of reasoning in multiagent systems.

- *Receiving new information*: new information can be represented as a propositional fact provided with a degree of plausibility;
- *evaluating new information*: the origin of new information decisively influences the agent's willingness to adopt it;
- *changing beliefs*: the agent uses belief revision techniques to change its epistemic state according to the new adopted information;
- *inference:* the agent's behavior is influenced by the most plausible beliefs resulting from its new epistemic state.

3.5 Conclusion

Trust plays an important role in many research areas of artificial intelligence, particularly in the semantic web and multiagent systems where the sources have to deal with conflicting information from other sources. Building on the socio-cognitive model of trust described in Castelfranchi and Falcone [90], and on previous work integrating trust and argumentation [310], in this chapter we presented a formal framework for modeling how different dimensions of the perceived trustworthiness of the source interact to determine the acceptability of the message, and how deviations from such expectation produce a specific feedback on source trustworthiness. Here, we applied this model to the case of sources exchanging and assessing arguments, but it could easily be extended to the exchange of any kind of factual information. The reason why we focused first on argumentation is because this provides a window on the agent's reasoning. The arguments in Section 3.2 are treated basically as black boxes, as it is most often the case in works based on abstract argumentation, in the vein of Dung [133]. This is significant in two respects. First, we did

not discuss the two-way relationship between source trustworthiness and trust in the message when what is being communicated is not the argument as a whole, but rather one of its constituents, e.g., a premise, its conclusion, or the inference rule licensing the argument, as in Parsons et al. [253]. Finding out that the source is mistaken on the truth of some premise (hence the argument is unsound) rather than on the truth of the inference (hence the argument is invalid) is likely to have very different effects for the feedback on the source, which will have to be investigated in future work. Second, we treat only the case of valid arguments, again as it is customary in abstract argumentation after Dung [133]. This is of course a huge idealization with respect to everyday argumentation: as underlined by Walton [318], we rarely exchange deductively valid arguments, while the vast majority of arguments are defeasible, which implies a different sort of consequence relation. Finally, the framework does not capture the cumulative effect of converging sources on argument acceptability. When more than one source offers the same information item, its acceptability is positively affected, as discussed in [88].

In this chapter, we have also justified and developed an approach to graded reinstatement in belief revision. The acceptability of arguments depends on a fuzzy labeling algorithm based on possibility theory. An agent will believe the conclusions of the accepted arguments, as well as their consequences. Arguments reinstatement induces the reinstatement of the beliefs grounded on these arguments. Arguments and beliefs are reinstated depending on the trustworthiness of the sources proposing them. The framework can be further improved following two directions: (i) specifying the trustworthiness degree by a cognitive model of trust and, (ii) embedding the proposed framework into an integrated one where also desires and, afterwards, goals are taken into account. The extent to which new information is accepted by an agent directly depends on the content of the new claim and on how much the agent trusts the source providing it. However, trust may also be influenced by information content. Indeed, even tough I might not particularly trust a source, if it provides me a claim which is consistent with my beliefs, I will not change my beliefs. However, my degree of trust for that source may increase. Trust depends thus on the agent's own beliefs in general and, in particular, on the agent's opinion about the capability of the source to convey useful information. In real-world situations, an agent's beliefs about a source may be incomplete, for they may derive from the opinions of other (partially) known agents and the agent may have had few (or none) exchanges with the source. On the other hand, only an agent endowed with goals and beliefs can trust another agent [89]. In other words, if an agent needs to trust a source, it is because it needs "something" from that source that could help it fulfill its own goals. Therefore, the agent's beliefs about the source's goals in comparison with its own goals must also play an important role in computing trust. These beliefs can be constructed from the agent's past interactions and the source's reputation and/or recommendations. Because we are aware that trust is not always the complement of distrust, here, we consider the bipolar side of trust. Our key idea here is to capture the fact that some pieces of information can contribute to increase or decrease trust, and other pieces of information can contribute to increase or decrease distrust.

Chapter 4

Argumentation for Explanation and Justification

4.1 Introduction

This chapter synthesizes my contribution about machine explanation and justification, where an argumentation-based module has been integrated into a question-answering system to support the user in a better understanding about the results provided by the system. These contribution have been published in a paper on the *Semantic Web* journal [74]: *Elena Cabrio, Serena Villata, Alessio Palmero Aprosio. A RADAR for information reconciliation in Question Answering systems over Linked Data. Semantic Web 8(4): 601-617 (2017).* These results benefited from the collaboration with Elena Cabrio (UNS) and Alessio Palmero Aprosio (FBK Trento). This line of work, having as goal the use of argumentation theory to explain machine decisions, is the core of the project with Accenture I coordinate, and that will start on July 2018. In this context, I will supervise the activity of the research engineer (Nicholas Halliwell) we recruited.

In the Web of Data, it is possible to retrieve heterogeneous information items concerning a single real-world object coming from different data sources, e.g., the results of a single SPARQL query on different endpoints. It is not always the case that these results are identical, it may happen that they conflict with each other, or they may be linked by some other relation like a specification. The automated detection of the kind of relationship holding between different instances of a single object with the goal of reconciling them is an open problem for consuming information in the Web of Data. In particular, this problem arises while querying the language-specific chapters of DBpedia [221]. Such chapters, well connected through Wikipedia instance interlinking, can in fact contain different information with respect to the English version. Assuming we wish to query a set of language-specific DBpedia SPARQL endpoints with the same query, the answers we collect can be either identical, or in some kind of specification relation, or they can be contradictory. Consider for instance the following example: we query a set of language-specific DBpedia chapters about *How tall is the soccer player Stefano Tacconi?*, receiving the following information: 1.88 from the Italian chapter and the German one, 1.93 from the French chapter, and 1.90 from the English one. How can I know what is the "correct" (or better, the more reliable) information, knowing that the height of a person is unique? Addressing such kind of issues is the goal of the present chapter. More precisely, in this chapter, we answer the research question: *how to reconcile information provided by the language-specific chapters of DBpedia?*

This open issue is particularly relevant to Question Answering (QA) systems over DBpedia [210], where

the user expects a unique (ideally correct) answer to her factual natural language question. A QA system querying different data sources needs to weight them in an appropriate way to evaluate the information items they provide accordingly. In this scenario, another open problem is how to explain and justify the answer the system provides to the user in such a way that the overall QA system appears transparent and, as a consequence, more reliable. Thus, our research question breaks down into the following subquestions:

1. How to automatically detect the relationships holding between information items returned by different language-specific chapters of DBpedia?
2. How to compute the reliability degree of such information items to provide a unique answer?
3. How to justify and explain the answer the QA system returns to the user?

First, we need to classify the relations connecting each piece of information to the others returned by the different data sources, i.e., the SPARQL endpoints of the language-specific DBpedia chapters. We adopt the categorization of the relations existing between different information items retrieved with a unique SPARQL query proposed by Cabrio et al. [76]. Up to our knowledge, this is the only available categorization that considers linguistic, fine-grained relations among the information items returned by language-specific DBpedia chapters, given a certain query. This categorization considers ten *positive* relations among heterogenous information items (referring to widely accepted linguistic categories in the literature), and three *negative* relations meaning inconsistency. Starting from this categorization, we propose the RADAR (ReconciliAtion of Dbpedia through ARgumentation) framework, that adopts a classification method to return the relation holding between two information items. This first step results in a graph-based representation of the result set where each information item is a node, and edges represent the identified relations.

Second, we adopt *argumentation theory* [132], a suitable technique for reasoning about conflicting information, to assess the acceptability degree of the information items, depending on the relation holding between them and the trustworthiness of their information source [112]. Roughly, an abstract argumentation framework is a directed labeled graph whose nodes are the arguments and the edges represent a *conflict* relation. Since positive relations among the arguments may hold as well, we rely on bipolar argumentation [92] that considers also a *positive* support relation.

Third, the graph of the result set obtained after the classification step, together with the acceptability degree of each information item obtained after the argumentation step, is used to justify and explain the resulting information ranking (i.e., the order in which the answers are returned to the user).

We evaluate our approach as standalone (i.e., over a set of heterogeneous values extracted from a set of language-specific DBpedia chapters), and through its integration in the QA system QAKiS [79], that exploits language-specific DBpedia chapters as RDF datasets to be queried using a natural language interface. The reconciliation module is embedded to provide a (possibly unique) answer whose acceptability degree is over a given threshold, and the graph structure linking the different answers highlights the underlying justification. Moreover, RADAR is applied to over 300 DBpedia properties in 15 languages, and the obtained resource of reconciled DBpedia language-specific chapters is released.

Even if information reconciliation is a way to enhance Linked Data quality, this chapter does not address the issue of Linked Data quality assessment and fusion [220, 63], nor ontology alignment. Finally, argumentation theory in this chapter is not exploited to find agreements over ontology alignments [283]. Note that our approach is intended to reconcile and explain the answer of the system to the user. Ontology alignment cannot be exploited to generate such a kind of explanations. This is why we decided to rely on argumentation theory that is a way to exchange and explain viewpoints. In this chapter, we have addressed

the open problem of reconciling and explaining a result set from language-specific DBpedia chapters using well known conflict detection and explanation techniques, i.e., argumentation theory.

We are not aware of any other available QA system that queries several information sources (in our case language-specific chapters of DBpedia) and then is able to verify the coherence of the proposed result set, and show *why* a certain answer has been provided. The merit of the present chapter is to describe the proposed framework (i.e., RADAR 2.0) with the addition of an extensive evaluation over standard QA datasets.

In the remainder of the chapter, Section 4.2 presents our reconciliation framework for language-specific DBpedia chapters, Section 4.3 reports on the experiments run over DBpedia to evaluate it, and Section 4.4 describes its integration in QAKiS. Section 4.5 reports on the related work. Finally, some conclusions are drawn.

4.2 RADAR 2.0: a Framework for Information Reconciliation

The RADAR 2.0 (ReconciliAtion of Dbpedia through ARgumentation) framework for information reconciliation is composed by three main modules (see Figure 4.1). It takes as input a collection of results from one SPARQL query raised against the SPARQL endpoints of the language-specific DBpedia chapters (Section 4.3 provides more details about the chapters considered in our experimental setting). Given such result set, RADAR retrieves two kinds of information: *(i)* the sources proposing each particular element of the result set, and *(ii)* the elements of the result set themselves. The first module of RADAR (module A, Figure 4.1) takes each information source, and following two different heuristics, assigns a *confidence degree* to the source. Such confidence degree will affect the reconciliation in particular with respect to the possible inconsistencies: information proposed by the more reliable source will obtain a higher acceptability degree. The second module of RADAR (module B, Figure 4.1) instead starts from the result set, and it matches every element with all the other returned elements, detecting the kind of relation holding between these two elements. The result of such module is a graph composed by the elements of the result set connected with each other by the relations of our categorization. Both the sources associated with a confidence score and the result set in the form of a graph are then provided to the third module of RADAR, the argumentation one (module C, Figure 4.1). The aim of such module is to reconcile the result set. The module considers all positive relations as a *support* relation and all negative relations as an *attack* relation, building a bipolar argumentation graph where each element of the result set is seen as an argument. Finally, adopting a bipolar fuzzy labeling algorithm relying on the confidence of the sources to decide the acceptability of the information, the module returns the acceptability degree of each argument, i.e., element of the result set. The output of the RADAR framework is twofold. First, it returns the acceptable elements (a threshold is adopted), and second the graph of the result set is provided, where each element is connected to the others by the identified relations (i.e., the explanation about the choice of the acceptable arguments returned).

In the remainder of this section, we will describe how the confidence score of the sources is computed (Section 4.2), and we will summarize the adopted categorization detailing how such relations are automatically extracted (Section 4.2). Finally, the argumentation module is described in Section 4.2.

Assigning a confidence score to the source

Language-specific DBpedia chapters can contain different information on particular topics, e.g. providing more or more specific information. Moreover, the knowledge of certain instances and the conceptualization of certain relations can be culturally biased. For instance, we expect to have more precise (and possibly more

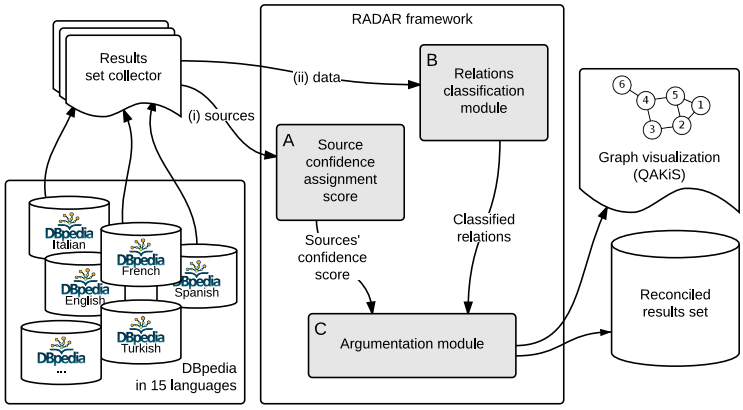


Figure 4.1: RADAR 2.0 framework architecture.

reliable) information on the Italian actor Antonio Albanese on the Italian DBpedia, than on the English or on the French ones.

To trust and reward the data sources, we need to calculate the reliability of the source with respect to the contained information items. In [77], an apriori confidence score is assigned to the endpoints according to their dimensions and solidity in terms of maintenance (the English chapter is assumed to be more reliable than the others on all values, but this is not always the case). RADAR 2.0 assigns, instead, a confidence score to the DBpedia language-specific chapter depending on the queried entity, according to the following two criteria:

- *Wikipedia page length*. The chapter of the longest language-specific Wikipedia page describing the queried entity is considered as fully trustworthy (i.e., it is assigned with a score of 1) while the others are considered less trustworthy (i.e., they are associated with a score < 1). In choosing such heuristic, we followed [51] that demonstrates that the article length is a very good predictor of its precision. The length is calculated on the Wikipedia dump of the considered language (# of characters in the text, ignoring image tags and tables). Thus, the longest page is assigned a score equal to 1, and a proportional score is assigned to the other chapters.

- *Entity geo-localization*. The chapter of the language spoken in the places linked to the page of the entity is considered as fully trustworthy (i.e., it is assigned with a score of 1) while the others are considered less trustworthy (i.e., they are associated with a score < 1). We assume that if an entity belongs to a certain place or is frequently referred to it, it is more likely that the DBpedia chapter of such country contains updated and reliable information. All Wikipedia page hyperlinks are considered, and their presence in GeoNames[1] is checked. If existing, the prevalent language in the place (following the GeoNames matching country-language[2]) is extracted, and to the corresponding chapter a score equal to 1 is assigned. As for page length, a proportional score is then assigned to the other chapters (i.e. if an entity has e.g. 10 links to places in Italy and 2 to places in Germany, the score assigned to the Italian DBpedia chapter is 1, while for the German chapter is 0.2).

[1] `http://www.geonames.org/`

[2] Such table connecting a country with its language can be found here: `http://download.geonames.org/export/dump/countryInfo.txt`.

Such metrics (the appropriateness of which for our purposes has been tested on the development set, see Section 4.3) are then summed and normalized with a score ranging from 0 to 1, where 0 is the least reliable chapter for a certain entity and 1 is the most reliable one. The obtained scores are then considered by the argumentation module (Section 4.2) for information reconciliation.

Relations classification

Cabrio et al. [76] propose a classification of the semantic relations holding among the different instances obtained by querying a set of language-specific DBpedia chapters with a certain query. More precisely, such categories correspond to the lexical and discourse relations holding among heterogeneous instances obtained querying two DBpedia chapters at a time, given a subject and an ontological property. In the following, we list the positive relations between values resulting from the data-driven study in [76]. Then, in parallel, we describe how RADAR 2.0 addresses the automatic classification of such relations.

Identity i.e., same value but in different languages (missing `owl:sameAs` link in DBpedia). E.g., `Dairy product` vs `Produits laitiers`

Acronym i.e., initial components in a phrase or a word. E.g., `PSDB` vs `Partito della Social Democrazia Brasiliana`

Disambiguated entity i.e., a value contains in the name the class of the entity. E.g., `Michael Lewis (Author)` vs `Michael Lewis`

Coreference i.e., an expression referring to another expression describing the same thing (in particular, non normalized expressions). E.g., `William Burroughs` vs `William S. Burroughs`

Given the high similarity among the relations belonging to these categories, we cluster them into a unique category called *surface variants* of the same entity. Given two entities, RADAR automatically detects the *surface variants* relation among them, if one of the following strategies is applicable: cross-lingual links[3], text identity (i.e. string matching), Wiki redirection and disambiguation pages.

Geo-specification i.e., ontological geographical knowledge. E.g., `Queensland` vs `Australia`

Renaming i.e., reformulation of the same entity name in time. E.g., `Edo`, old name of `Tokyo`

Given the way in which *renaming* has been defined in [76], it refers only to geographical renaming. For this reason, we merge it to the category *geo-specification*. RADAR classifies a relation among two entities as falling inside this category when in the GeoNames one entity is contained in the other one (*geo-specification* is a directional relation between two entities). We also consider the alternative names gazette included in GeoNames, and geographical information extracted from a set of English Wikipedia infoboxes, such as `Infobox former country`[4] or `Infobox settlement`.

[3]Based on WikiData, a free knowledge base that can be read and edited by humans and machines alike, `http://www.wikidata.org/`, where data entered in any language is immediately available in all other languages. In WikiData, each entity has the same ID in all languages for which a Wikipedia page exists, allowing us to overcome the problem of missing `owl:sameAs` links in DBpedia (that was an issue in DBpedia versions prior to 3.9). Moreover, WikiData is constantly updated (we use April 2014 release).

[4]For instance, we extract the property "today" connecting historical entity names with the current ones (reconcilable with GeoNames). We used Wikipedia dumps.

Meronymy i.e., a constituent part of, or a member of something. E.g., `Justicialist Party` is a part of `Front for Victory`

Hyponymy i.e., relation between a specific and a general word when the latter is implied by the former. E.g., `alluminio` vs `metal`

Metonymy i.e., a name of a thing/concept for that of the thing/concept meant. E.g., `Joseph Hanna` vs `Hanna-Barbera`

Identity:stage name i.e., pen/stage names pointing to the same entity. E.g., `Lemony Snicket` vs `Daniel Handler`

We cluster such semantic relations into a category called *inclusion*.[5] To detect this category of relations, RADAR exploits a set of features extracted from:

MusicBrainz[6] to detect when a musician plays in a band, and when a label is owned by a bigger label.

BNCF (Biblioteca Nazionale Centrale di Firenze) Thesaurus[7] for the broader term relation between common names.

DBpedia, in particular the datasets connecting Wikipedia, GeoNames and MusicBrainz trough the `owl:sameAs` relation.

WikiData for the *part of*, *subclass of* and *instance of* relations. It contains links to GeoNames, BNCF and MusicBrainz, integrating DBpedia `owl:sameAs`.

Wikipedia contains hierarchical information in: infoboxes (e.g. property `parent` for companies, `product` for goods, `alter ego` for biographies), categories (e.g., `Gibson guitars`), "see also" sections and links in the first sentence (e.g., *Skype was acquired by [United States]-based [Microsoft Corporation]*).

Inclusion is a directional relation between two entities (the rules we apply to detect *meronymy*, *hyponymy* and *stage name* allow us to track the direction of the relation, i.e. if $a \rightarrow b$, or $b \rightarrow a$).

Moreover, in the classification proposed in [76], the following negative relations (i.e., values mismatches) among possibly inconsistent data are identified:

Text mismatch i.e. unrelated entity. E.g., `Palermo` vs `Modene`

Date mismatch i.e. different date for the same event. E.g., `1215-04-25` vs `1214- 04-25`

Numerical mismatch i.e. different numerical values. E.g., `1.91` vs `1.8`

[5] Royo [274] defines both relations of *meronymy* and *hyponymy* as relations of *inclusion*, although they differ in the kind of inclusion defined (hyponymy is a relation of the kind "B is a type of A", while meronymy relates a whole with its different parts or members). Slightly extending Royo's definition, we joined to this category also the relation of *metonymy*, a figure of speech scarcely detectable by automatic systems due to its complexity (and *stage name*, that can be considered as a particular case of *metonymy*, i.e., the name of the character for the person herself).

[6] `http://musicbrainz.org/`

[7] `http://thes.bncf.firenze.sbn.it/`

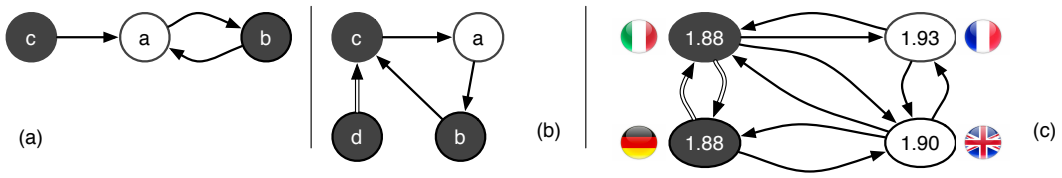


Figure 4.2: Example of (a) an AF, (b) a bipolar AF, and (c) example provided in the introduction modeled as a bipolar AF, where single lines represent attacks and double lines represent support.

RADAR labels a relation between instances (i.e., URIs) as negative, if every attempt to find one of the positive relations described above fails (i.e., negation as a failure). For numerical values, a *numerical mismatch* identifies different values.[8]

The reader may argue that a machine learning approach could have been applied to this task, but a supervised approach would have required an annotated dataset to learn the features. Unfortunately, at the moment there is no such training set available to the research community. Moreover, given the fact that our goal is to produce a resource as precise as possible for future reuse, the implementation of a rule-based approach allows us to tune RADAR to reward precision in our experiments, in order to accomplish our purpose.

Argumentation-based information reconciliation

This section details the RADAR 2.0 argumentation module, which extends the fuzzy labeling algorithm we introduced in Section 3.3.

Figure 4.2.a shows an example of an AF. The arguments are visualized as nodes of the argumentation graph, and the attack relation is visualized as edges. Gray arguments are the accepted ones. Using Dung's admissibility-based semantics [132], the set of accepted arguments is $\{b,c\}$.

Since we want to take into account the confidence associated with the information sources to compute the acceptability degree of arguments, we rely on the computation of fuzzy confidence-based degrees of acceptability. As the fuzzy labeling algorithm [112] exploits a scenario where the arguments are connected by an attack relation only, in Cabrio *et al.* [77] we have proposed a bipolar version of this algorithm, to consider also a positive, i.e., support, relation among the arguments (bipolar AFs) for the computation of the fuzzy labels of the arguments.

Let Ar be a fuzzy set of trustful arguments, and $Ar(A)$ be the membership degree of argument A in Ar, we have that $Ar(A)$ is given by the trust degree of the most reliable (i.e., trusted) source that offers argument A[9], and it is defined as follows: $Ar(A) = \max_{s \in \mathrm{src}(A)} \tau_s$ where τ_s is the degree to which source $s \in \mathrm{src}(A)$ is evaluated as reliable. The starting confidence degree associated with the sources is provided by RADAR's first module. The bipolar fuzzy labeling algorithm [77] assumes that the following two constraints hold: *(i)* an argument cannot attack and support another argument at the same time, and *(ii)* an argument cannot support an argument attacking it, and vice versa. These constraints underlie the construction of the bipolar AF itself. In the following, the attack relation is represented with $\rightarrow$, and the support relation with $\Rightarrow$.

[8] At the moment no tolerance is admitted, if e.g. the height of a person differs of few millimeters in two DBpedia chapters, the relation is labeled as *numerical mismatch*. We plan to add such tolerance for information reconciliation as future work.

[9] We follow the approach presented in Section 3.3 choosing the max operator ("optimistic" assignment of the labels), but the min operator may be preferred for a pessimistic assignment.

Table 4.1: *BAF*: $a \to b, b \to c, c \to a, d \Rightarrow c$

t	$\alpha_t(a)$	$\alpha_t(b)$	$\alpha_t(c)$	$\alpha_t(d)$
0	1	0.4	0.2	**1**
1	0.9	0.2	**0.6**	↓
2	0.65	0.15	↓	
3	0.52	0.25		
4	0.46	0.36		
5	0.43	**0.4**		
6	0.41	↓		
7	**0.4**			
8	↓			

Definition 61. Let $\langle Ar, \to, \Rightarrow \rangle$ be an abstract bipolar argumentation framework where Ar is a fuzzy set of (trustful) arguments, $\to \subseteq Ar \times Ar$ and $\Rightarrow \subseteq Ar \times Ar$ are two binary relations called attack and support, respectively. A bipolar fuzzy labeling is a total function $\alpha : Ar \to [0,1]$.

Such an α may also be regarded as (the membership function of) the fuzzy set of acceptable arguments where the label $\alpha(A) = 0$ means that the argument is outright unacceptable, and $\alpha(A) = 1$ means the argument is fully acceptable. All cases inbetween provide the degree of the acceptability of the arguments which may be considered accepted in the end, if they exceed a certain threshold.

A bipolar fuzzy labeling is defined as follows[10], where argument B is an argument attacking A and C is an argument supporting A:

Definition 62. (Bipolar Fuzzy Labeling) A total function $\alpha : Ar \to [0,1]$ is a bipolar fuzzy labeling iff, for all arguments A, $\alpha(A) = \text{avg}\{\min\{Ar(A), 1 - \max_{B:B \to A} \alpha(B)\}; \max_{C:C \Rightarrow A} \alpha(C)\}$.

When the argumentation module receives the elements of the result set linked by the appropriate relation and the confidence degree associated to each source, the bipolar fuzzy labeling algorithm is applied to the argumentation framework to obtain the acceptability degree of each argument. In case of cyclic graphs, the algorithm starts with the assignment of the trustworthiness degree of the source to the node, and then the value converges in a finite number of steps to the final label. Note that when the argumentation framework is composed by a cycle only, then all labels become equal to 0.5.

Consider the example in Figure 4.2.b, if we have $Ar(a) = Ar(d) = 1$, $Ar(b) = 0.4$ and $Ar(c) = 0.2$, then the fuzzy labeling algorithm returns the following labels: $\alpha(a) = \alpha(b) = 0.4$, $\alpha(c) = 0.6$, and $\alpha(d) = 1$. The step by step computation of the labels is shown in Table 4.1. Figure 4.2.c shows how the example provided in the introduction is modeled as a bipolar argumentation framework, where we expect the Italian DBpedia chapter to be the most reliable one, given that Stefano Tacconi is an Italian soccer player. The result returned by the bipolar argumentation framework is that the trusted answer is 1.88. A more precise instantiation of this example in the QA system is shown in the next section.

The fact that an argumentation framework can be used to provide an explanation and justify positions is witnessed by a number of applications in different contexts [35], like for instance practical reasoning [316], legal reasoning [36, 43], medical diagnosis [174]. This is the reason why we choose this formalism to reconcile information, compute the set of reliable information items, and finally justify this result. Other

[10]For more details about the bipolar fuzzy labeling algorithm, see Cabrio et al. [77].

possible solutions would be (weighted) voting mechanisms, where the preferences of some voters, i.e., the most reliable information sources, carry more weight than the preferences of other voters. However, voting mechanisms do not consider the presence of (positive and negative) relations among the items within the list, and no justification beyond the basic trustworthiness of the sources is provided to motivate the ranking of the information items.

Notice that argumentation is needed in our use case because we have to take into account the trustworthiness of the information sources, and it provides an explanation of the ranking, which is not possible with simple majority voting. Argumentation theory, used as a conflict detection technique, allows us to detect inconsistencies and consider the trustworthiness evaluation of the information sources, as well as proposing a single answer to the users. As far as we know, RADAR integrated in QAKiS is the first example of QA over Linked Data system coping with this problem and providing a solution. Simpler methods would not allow to cover both aspects mentioned above. We use bipolar argumentation instead of non-bipolar argumentation because we have not only the negative conflict relation but also the positive support relation among the elements of the result set.

4.3 RADAR experimental setting and evaluation

In this section, we describe the dataset on which we evaluate the RADAR framework (Section 4.3), and we discuss the obtained results (Section 4.3). Moreover, in Section 4.3 we describe the resource of reconciled DBpedia information we create and release.

Dataset

To evaluate the RADAR framework, we rely on the dataset presented in Cabrio et al. [76], the only available annotated dataset of possibly inconsistent information in DBpedia language-specific chapters to our knowledge. It is composed of 400 annotated pairs of values (extracted from English, French and Italian DBpedia chapters), a sample that is assumed to be representative of the linguistic relations holding between values in DBpedia chapters. Note that the size of the DBpedia chapter does not bias the type of relations identified among the values, nor their distribution, meaning that given a specific property, each DBpedia chapter deals with that property in the same way. We randomly divided such dataset into a development (to tune RADAR) and a test set, keeping the proportion among the distribution of categories.[11] Table 4.2 reports on the dataset statistics, and shows how many annotated relations belong to each of the categories (described in Section 4.2).

Table 4.2: Statistics on the dataset used for RADAR 2.0 evaluation

Dataset	*# triples*	*# annotated positive relations*			*# annotated negative relations*		
		Surface-form	Geo-specific.	Inclusion	Text mismatch	Date mismatch	Numerical mismatch
Dev set	104	28	18	20	13	13	12
Test set	295	84	48	55	36	37	35
Total	399	112	66	75	49	50	47

[11] The dataset is available at `http://www.airpedia.org/radar-1.0.nt.bz2`. The original work is based on DBpedia 3.9, but we updated it to DBpedia 2014. Thus, we deleted one pair, since the DBpedia page of one of the annotated entities does not exist anymore.

Results and discussion

Table 4.3 shows the results obtained by RADAR on the relation classification task on the test set. As baseline, we apply an algorithm exploiting only cross-lingual links (using WikiData), and exact string matching. Since we want to produce a resource as precise as possible for future reuse, RADAR has been tuned to reward precision (i.e., so that it does not generate false positives for a category), at the expense of recall (errors follow from the generation of false negatives for positive classes). As expected, the highest recall is obtained on the *surface form* category (our baseline performs even better than RADAR on such category). The *geo-specification* category has the lowest recall, either due to missing alignments between DBpedia and GeoNames (e.g. Ixelles and Bruxelles are not connected in GeoNames), or to the values complexity in the *renaming* subcategory (e.g., Paris vs First French Empire, or Harburg (quarter) vs Hambourg). In general, the results obtained are quite satisfying, fostering future work in this direction.

Table 4.3: Results of the system on relation classification

System	*Relation category*	*Precision*	*Recall*	F_1
RADAR 2.0	*surface form*	0.91	0.83	0.87
	geo-specification	0.94	0.60	0.73
	inclusion	0.86	0.69	0.77
	overall positive	1.00	0.74	0.85
	text mismatch	0.45	1	0.62
baseline	*surface form*	1.00	0.44	0.61
	geo-specification	0.00	0.00	0.00
	inclusion	0.00	0.00	0.00
	overall positive	1.00	0.21	0.35
	text mismatch	0.21	1	0.35

Since we consider *text mismatch* as a negative class (Section 4.2), it includes the cases in which RADAR fails to correctly classify a pair of values into one of the positive classes. For date and numerical mismatches, $F_1 = 1$ (detecting them is actually a trivial task, and therefore they are not included in Table 4.3. See footnote 8). *Overall positive* means that RADAR correctly understands the fact that the different answers to a certain query are all correct and not conflicting. RADAR precision in this case is 1, and it is important to underline this aspect in the evaluation, since this confirms the reliability of the released reconciled DBpedia in this respect. The overall positive result is higher than the partial results because in the precision of partial values we include the fact that if e.g., a *surface form* relation is wrongly labeled as *geo-specification*, we consider this mistake both as a false negative for *surface form*, and as a false positive for *geo-specification*. This means that RADAR is very precise in assigning positive relations, but it could provide a less precise classification into finer-grained categories.

Reconciled DBpedia resource

We applied RADAR 2.0 on 300 DBpedia properties - the most frequent in terms of chapters mapping such properties, corresponding to 47.8% of all properties in DBpedia. We considered ∼5M Wikipedia entities.

The outcoming resource, a sort of *universal DBpedia*, counts ∼50M of reconciled triples from 15 DBpedia chapters: Bulgarian, Catalan, Czech, German, English, Spanish, French, Hungarian, Indonesian, Italian, Dutch, Polish, Portuguese, Slovenian, Turkish. Notice that we did not consider the endpoint availability as a requirement to choose the DBpedia chapters: data are directly extracted from the resource.

For functional properties, the RADAR framework is applied as described in Section 4.2. In contrast, the strategy to reconcile the values of non-functional properties is slightly different: when a list of values is admitted (e.g. for properties `child` or `instruments`), RADAR merges the list of the elements provided by the DBpedia chapters, and ranks them with respect to the confidence assigned to their source, after reconciling positive relations only (there is no way for lists to understand if an element is incorrect or just missing, e.g. in the list of the instruments played by John Lennon). But since the distinction between functional/non-functional properties is not precise in DBpedia, we manually annotated the 300 properties with respect to this classification, to allow RADAR to apply the correct reconciliation strategy, and to produce a reliable resource. In total, we reconciled 3.2 million functional property values, with an average accuracy computed from the precision and recall reported in Table 4.3. This resource is available here: `http://qakis.org/resources.htm`.

Moreover, we carried out a merge and a light-weight reconciliation of DBpedia classes applying the strategy called "DBpedia CL" in [6] where "CL" stands for cross-language (e.g., *Michael Jackson* is classified as a `Person` in the Italian and German DBpedia chapters, an `Artist` in the English DBpedia and a `MusicalArtist` in the Spanish DBpedia. As `Person`, `Artist` and `MusicalArtist` lie on the same path from the root of the DBpedia ontology, all of them are kept and used to classify *Michael Jackson*.

4.4 Integrating RADAR in a QA system

We integrate RADAR into a QA system over language-specific DBpedia chapters, given the importance that information reconciliation has in this context. Indeed, a user expects a unique (and possibly correct) answer to her factual natural language question, and would not trust a system providing her with different and possibly inconsistent answers coming out of a black box. A QA system querying different data sources needs therefore to weight in an appropriate way such sources in order to evaluate the information items they provide accordingly.

As QA system we selected QAKiS (Question Answering wiKiFramework-based System) [79], because it allows *i)* to query a set of language-specific DBpedia chapters using a natural language interface, and *ii)* its modular architecture can be flexibly modified to account for the proposed extension. QAKiS addresses the task of QA over structured knowledge-bases (e.g., DBpedia) [69], but taking into account also unstructured relevant information, e.g., Wikipedia pages. It implements a relation-based match for question interpretation, to convert the user question into a query expressed in a query language (e.g., SPARQL), making use of relational patterns (automatically extracted from Wikipedia), that capture different ways to express a certain relation in a language. The actual version of QAKiS targets questions containing a Named Entity (NE) related to the answer through one property of the ontology, as *Which river does the Brooklyn Bridge cross?*. Such questions match a single pattern.

In QAKiS, the SPARQL query created after the question interpretation phase is sent to the SPARQL endpoints of the language-specific DBpedia chapters (i.e., English, French, German and Italian) for answer retrieval. The set of retrieved answers from each endpoint is then sent to RADAR 2.0 for answer reconciliation.

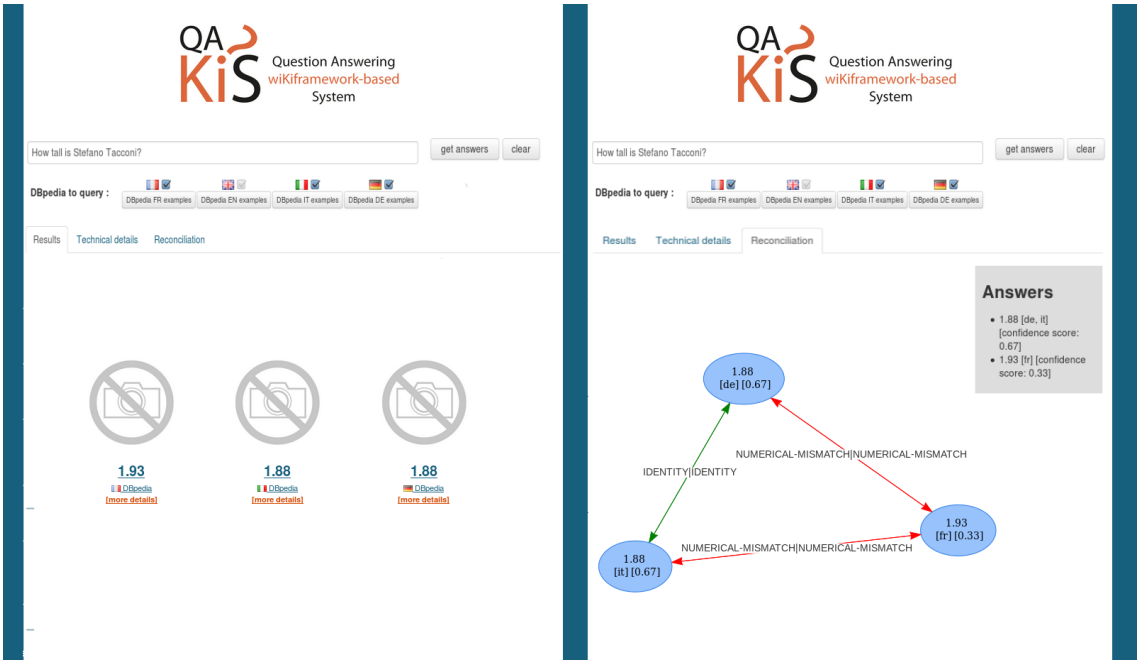


Figure 4.3: QAKiS + RADAR demo (functional properties)

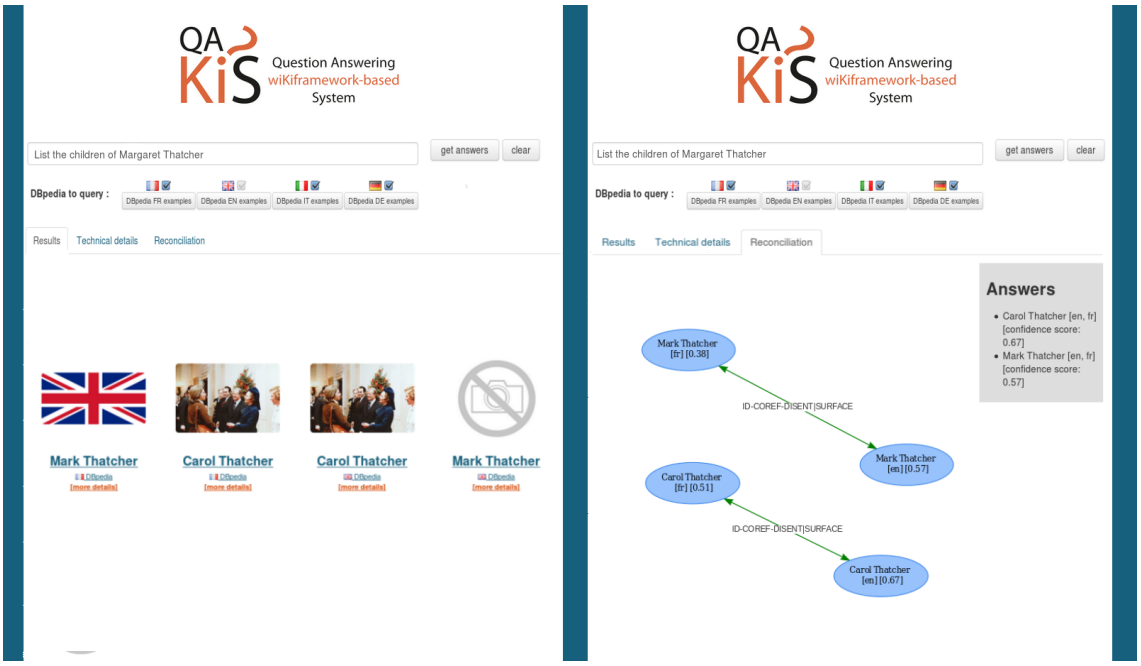


Figure 4.4: QAKiS + RADAR demo (non-functional properties)

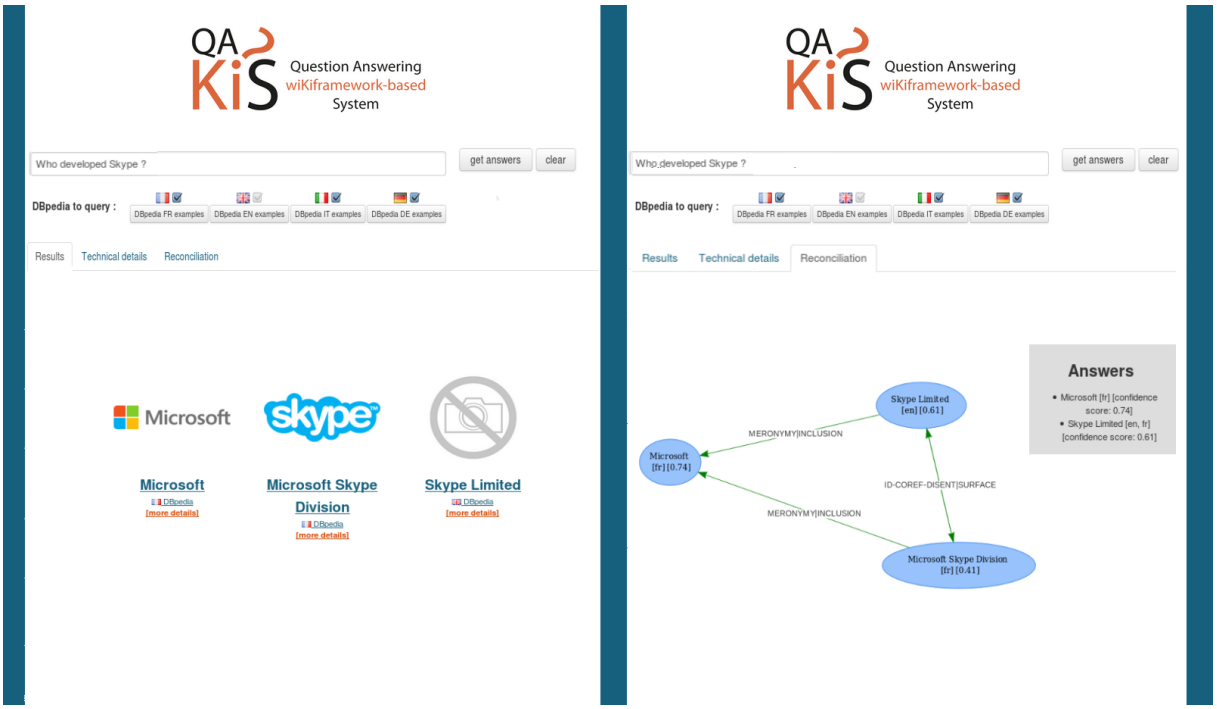


Figure 4.5: Example about the question *Who developed Skype?*

To test RADAR integration into QAKiS[12], the user can select the DBpedia chapter she wants to query besides English (that must be selected as it is needed for NE recognition), i.e., French, German or Italian DBpedia. Then the user can either write a question or select among a list of examples. Clicking on the tab *Reconciliation*, a graph with the answers provided by the different endpoints and the relations among them is shown to the user (as shown in Figures 4.3 and 4.4 for the questions *How tall is Stefano Tacconi?*, and *List the children of Margaret Thatcher*, respectively). Each node has an associated confidence score, resulting from the fuzzy labeling algorithm (described in Section 4.2). Moreover, each node is related to the others by a relation of support or attack, and a further specification of such relations according to the categories described in Section 4.2 is provided to the user as answer justification of why the information items have been reconciled and ranked in this way.

Looking at these examples, the reader may argue that the former question can be answered by a simple majority voting (Figure 4.3), and the latter can be answered by a grouping based on surface forms (Figure 4.4), without the need to introduce the complexity of the argumentation machinery. However, if we consider the following example from our dataset, the advantage of using argumentation theory becomes clear. Let us consider the question *Who developed Skype?*: in this case, we retrieve three different answers, namely Microsoft (from FR DBpedia), Microsoft Skype Division (from FR DBpedia), and Skype Limited (EN DBpedia). The relations assigned by RADAR are visualized in Figure 4.5. The answer, with the associated weights, returns first Microsoft (FR) with a confidence score of 0.74, and second, Skype Limited (EN, FR) with a confidence score of 0.61. Note that this result cannot be achieved with simple majority voting nor with grouping based on surface forms.

[12]Demo at `http://qakis.org`

QA experimental setting

To provide a quantitative evaluation of RADAR integration into QAKiS on a standard dataset of natural language questions, we consider the questions provided by the organizers of the QALD challenge (Question Answering over Linked Data challenge), now at its fifth edition, for the DBpedia track.[13] More specifically, we collect the questions sets of QALD-2 (i.e. 100 questions of the training and 100 questions of the test sets), the test set of QALD-4 (i.e. 50 questions), and the questions sets of QALD-5 (50 additional training questions with respect to the previous years training set, and 59 questions in the test sets). These 359 questions correspond to all the questions released in the five years of the QALD challenge (given the fact that the questions of QALD-1 are included into the question set of QALD-2, and the question set of QALD-3 is the same as QALD-2, but translated into 6 languages, and the training sets of QALD-4 and 5 include all the questions of QALD-2). QALD-3 also provides natural language questions for Spanish DBpedia, but given that the current version of QAKiS cannot query the Spanish DBpedia, we could not use this question set.

We extract from this reference dataset of 359 questions, the questions that the current version of QAKiS is built to address (i.e. questions containing a NE related to the answer through one property of the ontology), corresponding to 26 questions in QALD-2 training set, 32 questions in QALD-2 test sets, 12 in QALD-4 test set, 18 in QALD-5 training set, and 11 in QALD-5 test set. The discarded questions require either some form of aggregation (e.g., counting or ordering), information from datasets different than DBpedia, involve *n*-ary relations, or are boolean questions. We consider these 99 questions as the QALD reference dataset for our experiments.

Results on QALD answers reconciliation

We run the questions contained into our QALD reference dataset on the English, German, French and Italian chapters of DBpedia. Since the questions of QALD were created to query the English chapter of DBpedia only, it turned out that only in 43/99 cases at least two endpoints provide an answer (in all the other cases the answer is provided by the English chapter only, not useful for our purposes). For instance, given the question *Who developed Skype?* the English DBpedia provides *Skype Limited* as the answer, while the French one outputs *Microsoft* and *Microsoft Skype Division*. Or given the question *How many employees does IBM have?*, the English and the German DBpedia chapters provide 426751 as answer, while the French DBpedia 433362. Table 4.5 lists these 43 QALD questions, specifying which DBpedia chapters (among the English, German, French and Italian ones) contain at least one value for the queried relation. This list of question is the reference question set for our evaluation.

We evaluated the ability of RADAR 2.0 to correctly classify the relations among the information items provided by the different language-specific SPARQL endpoints as answer to the same query, w.r.t. a manually annotated goldstandard, built following the methodology in Cabrio et al. [76]. More specifically, we evaluate RADAR with two sets of experiments: in the first case, we start from the answers provided by the different DBpedia endpoints to the 43 QALD questions, and we run RADAR on it. In the second case, we add QAKiS in the loop, meaning that the data we use as input for the argumentation module are directly produced by the system. In this second case, the input are the 43 natural language questions.

Table 4.4 reports on the results we obtained for the two experiments. As already noticed before, the QALD dataset was created to query the English chapter of DBpedia only, and therefore this small

[13] http://www.sc.cit-ec.uni-bielefeld.de/qald/

dataset does not capture the variability of possibly inconsistent answers that can be found among DBpedia language-specific chapters. Only three categories of relations are present in this data – *surface forms*, *geo-specification*, and *inclusion* – and for this reason RADAR has outstanding performances on it when applied on the correct mapping between NL questions and the SPARQL queries. When QAKiS is added into the loop, its mistakes in interpreting the NL question and translating it into the correct SPARQL query are propagated in RADAR (that receives in those cases a wrong input), decreasing the total performances.

Notice that in some cases the question interpretation can be tricky, and can somehow bias the evaluation of the answers provided by the system. For instance, for the question *Which pope succeeded John Paul II?*, the English DBpedia provides *Benedict XVI* as the answer, while the Italian DBpedia provides also other names of people that were successors of John Paul II in other roles, as for instance in being the Archbishop of Krakow. But since in the goldstandard this question is interpreted as being the successor of John Paul II in the role of Pope, only the entity *Benedict XVI* is accepted as correct answer.

When integrated into QAKiS, RADAR 2.0 outperforms the results obtained by a preliminary version of the argumentation module, i.e. RADAR 1.0 [77], for the positive relation classification (the results of the argumentation module only cannot be strictly compared with the results obtained by RADAR 2.0, since *i)* in its previous version the relation categories are different and less fine-grained, and *ii)* in [77] only questions from QALD-2 were used in the evaluation), showing an increased precision and robustness of our framework. Note that this evaluation is not meant to show that QAKiS performance is improved by RADAR. Actually, RADAR does not affect the capacity of QAKiS to answer questions: RADAR is used to disambiguate among multiple answers retrieved by QAKiS in order to provide to the user the most reliable (and hopefully correct) one.

One of the reasons why RADAR is implemented as a framework that can be integrated on top of an existing QA system architecture (and is therefore system-independent), is because we would like it to be tested and exploited by potentially all QA systems querying more than one DBpedia chapter (up to our knowledge QAKiS is the only one at the moment, but given the potential increase in the coverage of a QA system querying multiple DBpedia language-specific chapters [69], we expect other systems to take advantage of these interconnected resources soon).

4.5 Related work

The present chapter is an extended version of our previous work [78, 77, 68] introducing RADAR 1.0. The following common points are present: the idea of using argumentation theory to detect inconsistencies over the result set of a question answering system exploiting DBpedia, and the bipolar extension of the original fuzzy labeling algorithm [112] to judge an argument's acceptability in presence of both support and attack relations. However, the present chapter presents a substantial extension with respect to this preliminary work. More specifically, the main enhancements are reported in the following:

Relation categorization. RADAR 2.0 exploits the categorization we introduced in [76], as mentioned in Section 4.2. However, the work presented in [76] is purely theoretic and the contribution here is to study how to make RADAR 2.0 match these linguistic relations with respect to the DBpedia use case. Moreover, the categorization of the possible relations holding between the information items we adopt here is different (more linguistically-motivated) and more fine-grained than the one we used in [77]. This fine-grained categorization allows for a more insightful justification graph.

Table 4.4: Results on QALD relation classification

System	*Relation category*	*Precision*	*Recall*	F_1
RADAR 2.0 (only)	*surface form*	1.00	0.98	0.99
	geo-specification	0.88	0.80	0.84
	inclusion	0.80	1.00	0.88
	overall positive	1.00	0.98	0.99
baseline	*surface form*	1.00	0.97	0.98
	geo-specification	0.00	0.00	0.00
	inclusion	0.00	0.00	0.00
	overall positive	1.00	0.86	0.92
QAKiS + RADAR 2.0	*surface form*	1.00	0.59	0.74
	geo-specification	0.88	0.80	0.84
	inclusion	0.80	1.00	0.88
	overall positive	1.00	0.63	0.77
QAKiS + baseline	*surface form*	1.00	0.58	0.74
	geo-specification	0.00	0.00	0.00
	inclusion	0.00	0.00	0.00
	overall positive	1.00	0.52	0.68
QAKiS + RADAR 1.0 [77] (on QALD-2 questions only)	**overall positive**	0.54	0.56	0.55

Relation extraction. The relations holding between the elements of the result set are here automatically extracted with the application of more robust techniques than in [77]. More precisely, the way RADAR 2.0 extracts these relations in an automated way is different from the way RADAR 1.0 extracts them: RADAR 2.0 adopts external resources to improve the extraction of the correct relation, such as MusicBrainz, the BNCF (Biblioteca Nazionale Centrale di Firenze), DBpedia and Wikipedia, GeoNames, and WikiData.

Evaluation. While in [68] only data from QALD-2 has been used, here we use all data available from the QALD challenges (all editions), and the Italian chapter of DBpedia is added as RDF dataset to be queried with QAKiS (not present in our previous works on the topic). Moreover, the results presented in this chapter show a higher precision with respect to the results obtained with RADAR 1.0 and reported in [77] (F_1 increments from 0.55 to 0.77 for the positive relation classification if we consider QALD-2 data only). In addition, the new evaluation considers 15 DBpedia chapters instead of the 3 chapters used in [68], i.e., English, German and French.

Resource. Differently from [77] where no resource resulted from the inconsistencies detection process, here we generate a resource applying the proposed framework to 15 reconciled language-specific DBpedia chapters, and we release it.

State-of-the-art QA systems over Linked Data generally address the issue of question interpretation mapping a natural language question to a triple-based representation (see [210] for an overview). Moreover, they examine the potential of open user friendly interfaces for the Semantic Web to support end users in reusing and querying the Semantic Web content. None of these systems considers language-specific DBpedia chapters, and they do not provide a mechanism to reconcile the different answers returned by heterogenous endpoints. Finally, none of them provides explanations about the answer returned to the user.

Several works address alignment agreement based on argumentation theory. More precisely, Laera et al. [190] address alignment agreement relying on argumentation to deal with the arguments which attack or support the candidate correspondences among ontologies. Doran et al. [128] propose a methodology to identify subparts of ontologies which are evaluated as sufficient for reaching an agreement, before the argumentation step takes place, and dos Santos and Euzenat [283] present a model for detecting inconsistencies in the selected sets of correspondences relating ontologies. In particular, the model detects logical and argumentation inconsistency to avoid inconsistencies in the agreed alignment. We share with these approaches the use of argumentation to detect inconsistencies, but RADAR goes beyond them: we identify in an automated way relations among information items that are more complex than `owl:sameAs` links (as in ontology alignment). Moreover, these approaches do not consider trust-based acceptance degrees of the arguments, lacking to take into account a fundamental component in the arguments' evaluation, namely their sources.

We mentioned these works applying argumentation theory to address ontology alignment agreements as examples of applications of this theory to open problems in the Semantic Web domain. Actually, the two performances cannot be compared to show the superiority of one of the two approaches, as the task is different.

4.6 Conclusion

In this chapter, we have introduced and evaluated the RADAR 2.0 framework for information reconciliation over language-specific DBpedia chapters. The framework is composed of three main modules: a module computing the confidence score of the sources depending either on the length of the related Wikipedia page or on the geographical characterization of the queried entity, a module retrieving the relations holding among the elements of the result set, and finally a module computing the reliability degree of such elements depending on the confidence assigned to the sources and the relations among them. This third module is based on bipolar argumentation theory, and a bipolar fuzzy labeling algorithm [77] is exploited to return the acceptability degrees. The resulting graph of the result set, together with the acceptability degrees assigned to each information item, justifies to the user the returned answer and it is the result of the reconciliation process. The evaluation of the framework shows the feasibility of the proposed approach. Moreover, the framework has been integrated in the question answering system over Linked Data called QAKiS, allowing to reconcile and justify the answers obtained from four language-specific DBpedia chapters (i.e. English, French, German and Italian). Finally, the resource generated applying RADAR to 300 properties in 15 DBpedia chapters to reconcile their values is released.

There are several points to be addressed as future work. First, the user evaluation should not be underestimated: we will soon perform an evaluation to verify whether our answer justification in QAKiS appropriately suits the needs of the data consumers, and to receive feedback on how to improve such visualization. Second, at the present stage we assign a confidence score to each source following two criteria, however another possibility is to let the data consumer itself assign such confidence degree to the sources

depending on the kind of information she is looking for. Finally, the proposed framework is not limited to the case of multilingual chapters of DBpedia. The general approach RADAR is based on allows to extend it to various cases like inconsistent information from multiple English data endpoints. The general framework would be the same, the only part to be defined are the rules to extract the relations among the retrieved results. Investigating how a module of this type can be adopted as a fact checking module is part of our future research plan.

Table 4.5: QALD questions used in the evaluation (in bold the ones correctly answered by QAKiS; *x* means that the corresponding language specific DBpedia chapter (EN, FR, DE, IT) contains at least one value for the queried relation; *dbo* means DBpedia ontology)

ID, question set	*Question*	*DBpedia relation*	*EN*	*FR*	*DE*	*IT*
84, QALD-2 train	Give me all movies with Tom Cruise.	`starring`	x	x	x	
10, QALD-2 train	**In which country does the Nile start?**	`sourceCountry`	x	x		
63, QALD-2 train	**Give me all actors starring in Batman Begins.**	`starring`	x	x	x	x
43, QALD-2 train	Who is the mayor of New York City?	`leaderName`	x		x	x
54, QALD-2 train	Who was the wife of U.S. president Lincoln?	`spouse`	x	x		
6, QALD-2 train	**Where did Abraham Lincoln die?**	`deathPlace`	x	x	x	
31, QALD-2 train	**What is the currency of the Czech Republic?**	`currency`	x	x	x	x
73, QALD-2 train	**Who owns Aldi?**	`keyPerson`	x	x		x
20, QALD-2 train	**How many employees does IBM have?**	`numberOfEmployees`	x	x	x	x
33, QALD-2 train	**What is the area code of Berlin?**	`areaCode`	x			
2, QALD-2 test	**Who was the successor of John F. Kennedy?**	`successor`	x	x		
4, QALD-2 test	How many students does the Free University in Amsterdam have?	`numberOfStudents`	x	x	x	
14, QALD-2 test	Give me all members of Prodigy.	`bandMember`	x	x		
20, QALD-2 test	**How tall is Michael Jordan?**	`height`	x		x	x
21, QALD-2 test	**What is the capital of Canada?**	`capital`	x	x	x	x
35, QALD-2 test	**Who developed Skype?**	`product`	x	x		
38, QALD-2 test	How many inhabitants does Maribor have?	`populationTotal`	x			x
41, QALD-2 test	**Who founded Intel?**	`foundedBy`	x	x		x
65, QALD-2 test	**Which instruments did John Lennon play?**	`instrument`	x	x		
68, QALD-2 test	**How many employees does Google have?**	`numberOfEmployees`	x	x		x
74, QALD-2 test	**When did Michael Jackson die?**	`deathDate`	x	x	x	
76, QALD-2 test	**List the children of Margaret Thatcher.**	`child`	x	x		
83, QALD-2 test	How high is the Mount Everest?	`elevation`	x	x		x
86, QALD-2 test	**What is the largest city in Australia?**	`largestCity`	x	x		
87, QALD-2 test	Who composed the music for Harold and Maude?	`musicComposer`	x		x	x
34, QALD-4 test	**Who was the first to climb Mount Everest?**	`firstAscentPerson`	x		x	
21, QALD-4 test	**Where was Bach born?**	`birthPlace`	x	x	x	x
32, QALD-4 test	**In which countries can you pay using the West African CFA franc?**	`currency`	x		x	
12, QALD-4 test	How many pages does War and Peace have?	`numberOfPages`	x	x		
36, QALD-4 test	Which pope succeeded John Paul II?	`successor`	x			x
30, QALD-4 test	When is Halloween?	`date`	x	x		
259, QALD-5 train	Who wrote The Hunger Games?	`author`	x	x		
280, QALD-5 train	**What is the total population of Melbourne, Florida?**	`populationTotal`	x	x		x
282, QALD-5 train	In which year was Rachel Stevens born?	`birthYear`	x	x	x	x
283, QALD-5 train	**Where was JFK assassinated?**	`deathPlace`	x	x	x	x
291, QALD-5 train	**Who was influenced by Socrates?**	`influencedBy`	x	x		
295, QALD-5 train	Who was married to president Chirac?	`spouse`	x	x		
298, QALD-5 train	**Where did Hillel Slovak die?**	`deathPlace`	x	x	x	x
7, QALD-5 test	Which programming languages were influenced by Perl?	`influencedBy`	x	x	x	x
18, QALD-5 test	**Who is the manager of Real Madrid?**	`manager`	x	x		
19, QALD-5 test	**Give me the currency of China.**	`country`	x		x	
32, QALD-5 test	What does the abbreviation FIFA stand for?	`name`	x		x	x
47, QALD-5 test	**Who were the parents of Queen Victoria?**	`parent`	x		x	x

Chapter 5

Mining natural language argumentation

5.1 Introduction

This chapter synthesizes my contributions in the area of argument mining, dealing with the detection of the argument components and the relations holding among them from raw texts. These contributions are across the Natural Language Processing research area and the computational models of argument one. These contributions have been published in several venues:

- *Elena Cabrio, Serena Villata. Natural Language Arguments: A Combined Approach. 20th European Conference on Artificial Intelligence (ECAI 2012): 205-210* [72],
- *Elena Cabrio, Serena Villata. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012): 208-212* [71],
- *Elena Cabrio, Serena Villata, Fabien Gandon. A Support Framework for Argumentative Discussions Management in the Web. 10th International Conference on Semantic Web: 412-426* [75],
- *Tom Bosc, Elena Cabrio, Serena Villata. DART: a Dataset of Arguments and their Relations on Twitter. Tenth International Conference on Language Resources and Evaluation (LREC 2016)* [60],
- *Tom Bosc, Elena Cabrio, Serena Villata. Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media. Computational Models of Argument (COMMA 2016): 21-32* [59],
- *Mihai Dusmanu, Elena Cabrio, Serena Villata. Argument Mining on Twitter: Arguments, Facts and Sources. 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017): 2317-2322* [136],
- *Stefano Menini, Elena Cabrio, Sara Tonelli, Serena Villata. Never Retreat, Never Retract: Argumentation Analysis for Political Speeches. Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)* [223],
- *Elena Cabrio, Serena Villata. A natural language bipolar argumentation approach to support users in online debate interactions. Argument & Computation 4(3): 209-230 (2013)* [70].

The contributions reported in this chapter are the results of several collaborations: first of all, the collaboration with Elena Cabrio (UNS), with whom we initiated this research field and with whom I have a fruitful collaboration since 2012; second, the collaboration with Sara Tonelli and Stefano Menini (FBK Trento). In the context of these contributions, I have supervised the activity of one research engineer (Tom Bosc), and one internship of 3 months (Mihai Dusmanu).

If you had the dream that one day, in the broad Artificial Intelligence (AI) area, Natural Language Processing (NLP) researchers and Knowledge Representation and Reasoning (KRR) researchers were able to sit down together at the table of a joint panel, discussing on how to make progress and realize automated argument detection, then this chapter is for you. This is the story of a research area called *Argument Mining* (AM), and how it has become an important topic in AI.

The first approaches to what is now called argument mining started to appear around 2010, when the first methods to mine (different connotations of) *arguments* from natural language documents were proposed: [299] introduced the definition of argumentative zoning for scientific articles, and [227] proposed a way to detect arguments from legal texts. Since these seminal approaches, the need for automated methods to mine arguments and the relations among them from natural language text was brought to light, but it was only briefly touched upon. The parallel advances, from the formal point of view in the research field of computational models of arguments, and from the point of view of the computational techniques for learning and understanding human language content in the NLP and the Machine Learning fields, boosted the almost contemporary organization of two events in 2014 targeting open discussions about the challenge of mining arguments from text. Both the workshop on Argument Mining[1] co-located with ACL, and the workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing [2] we organized, shared the same goal: bringing together the communities of NLP and of formal argumentation to jointly work towards the definition of the new research area of *argument mining*. Since then, two Dagstuhl Seminars have been organized on such topic[3], the Argument Mining workshop holds every year, two tutorials on AM have been given at IJCAI-2016[4] and ACL-2016[5], three ESSLLI courses[6] in 2017, and AM has became a topic in major AI and NLP conferences. All these clues prove its growing importance in AI.

Argument mining involves several research areas from the AI panorama: NLP provides the methods to process natural language text, to identify the arguments and their components (i.e., premises and claims) in texts and to predict the relations among such arguments, KRR contributes with the reasoning capabilities upon the retrieved arguments and relations so that, for instance, fallacies and inconsistencies can be automatically identified in such texts, and Human-Computer Interaction guides the design of good human-computer digital argument-based supportive tools.

This chapter is organized as follows: Section 5.2 presents the argument mining framework and its main tasks, and then I detail the different scenarios and challenges I addressed in the argument mining area, namely online debate platforms (Section 5.3), the Wikipedia revision history (Section 5.4), social media with a particular attention to Twitter data (Section 5.5), and political speeches (Section 5.6). Related work compare the proposed approaches to the existing literature, and a discussion about strong and weak points of the raising argument mining field concludes the chapter.

[1] https://goo.gl/kF4Eep

[2] https://goo.gl/ttVUZk

[3] I.e., Debating Technologies (https://goo.gl/osqEY3) and Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments (https://goo.gl/jS1Co6)

[4] https://goo.gl/kd4456

[5] http://acl2016tutorial.arg.tech/

[6] https://goo.gl/Cw1FLC

5.2 Background

Argument(ation) mining has been defined as "the general task of analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and automatically analyze the data at hand" [162]. Two stages are crucial in the argument mining framework:

Arguments' extraction : The first stage is the identification of arguments within the input natural language text. This step may be further split in two different stages such as the detection of argument components (e.g., claim, premises) and the further identification of their textual boundaries. Many approaches have recently been proposed to address such task, that adopt different methods like Support Vector Machines (SVM) [248, 204, 289, 235, 12], Naïve Bayes classifiers [137], Logistic Regression [198].

Relations' prediction : The second stage consists in predicting what are the relations holding between the arguments identified in the first stage. This is an extremely complex task, as it involves high-level knowledge representation and reasoning issues. The relations between the arguments may be of heterogeneous nature, like *attacks* and *supports*. They are used to build the argument graphs, in which the relations connecting the retrieved arguments (i.e., the nodes in the graph) correspond to the edges. Different methods have been employed to address this task, from standard SVMs to Textual Entailment [70]. This stage is also in charge of predicting, in structured argumentation, the internal relations of the argument's components, such as the connection between the premises and the claim [289].

To clarify such tasks, let us consider the following example from the political debate of the Campaign "Trump - Clinton" on September 2016.[7] The first task of the argument mining framework consists in detecting the arguments from the text. In the example below, we highlight the arguments that can be identified (premises underlined and claims in bold):

A_1: *She talks about solar panels. We invested in a solar company, our country.* ***That was a disaster****.* *They lost plenty of money on that one. Now, look, I'm a great believer in all forms of energy, but we're putting a lot of people out of work.*

A_2: *Well,* ***I'm really calling for major jobs*** *because the wealthy are going create tremendous jobs. They're going to expand their companies. They're going to do a tremendous job.*

It appears evident that the argumentative sentences "in the wild", i.e., in natural language text as the ones reported in the examples, are pretty far from the prototypical argumentation patterns usually investigated in KRR, increasing the complexity of the task.

Let us consider now another example from an online debate about *Random sobriety tests for drivers*[8], where we identify again premises and claims.

A_3: Little evidence random alcohol tests deter drunk driving. *There is a dearth of research regarding the deterrent effect of checkpoints. The only formally documented research regarding deterrence is a survey of Maryland's "Checkpoint Strikeforce" program. The survey found no deterrent effect:* **To date, there is no evidence to indicate that this campaign, which involves a number of sobriety checkpoints and media**

[7] Debate extracted from the Commission on Presidential Debates (`http://debates.org`).

[8] `http://www.debatepedia.com/en/index.php/Debate:_Random_sobriety_tests_for_drivers`

activities to promote these efforts, has had any impact on public perceptions, driver behaviors, or alcohol-related motor vehicle crashes and injuries. This conclusion is drawn after examining statistics for alcohol-related crashes, police citations for impaired driving, and public perceptions of alcohol-impaired driving risk.

A_4: **Random breath testing doesn't necessarily lower drunk driving.** Many countries have had random testing for some time and have seen no real fall in drink driving figures.

A_5: **Random sobriety tests for drivers are effective at deterring drunk driving.**

Given these three arguments, the relations among them have to be predicted. Let us consider that the two relations we aim at identifying are *attack* (a negative relation between two arguments, e.g., a contradiction) and *support* (a positive relation between two arguments) only. In this case, we have that argument A_3 supports argument A_4, and argument A_4 attacks argument A_5.

It is important to underline at this point that argument mining differs from well known *opinion mining* (or *sentiment analysis*): while opinion mining focuses on understanding *what* users think about a certain topic or product, argument mining revolves around *why* users have a certain opinion about a topic or product.

Both the main argument mining tasks require high-quality annotated corpora to train and to evaluate the performances of automated approaches. The reliability of an annotated corpus is guaranteed by the calculation of the inter-annotator agreement that measures the degree of agreement in performing the annotation task among the involved annotators. For instance, when building a dataset for relation prediction, the statistical measure to be used to calculate the inter-rater agreement among the labels assigned by the annotators is the Cohen's kappa coefficient which takes into account also agreement occurring by chance. The equation for κ is $\kappa = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)}$ where $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the hypothetical probability of chance agreement. If the raters are in complete agreement then $\kappa = 1$, if there is no agreement among the raters other than what would be expected by chance, $\kappa = 0$. For NLP tasks, the agreement is considered as significant when $\kappa > 0.6$.[9]

5.3 A natural language bipolar argumentation approach for online debate interactions

In the last years, the Web has changed in the so called Social Web. The Social Web has seen an increasing number of applications like Twitter[10], Debatepedia[11], Facebook[12] and many others, which allow people to express their opinions about different issues. Let us consider for instance the following debate published on Debatepedia: the issue of the debate is "Making Internet a right only benefits society". The participants have proposed various pro and con arguments concerning this issue, e.g., a pro argument claims that the Internet delivers freedom of speech, and a con argument claims that the Internet is not as important as real rights like the freedom from slavery. These kinds of debates are composed by tens of arguments in favour or against a proposed issue. The main difficulty for newcomers is to understand the current holding position in the debate, i.e., to understand which are the arguments that are accepted at a certain moment. This difficulty is twofold: first, the participants have to remember all the different, possibly long, arguments and understand

[9] For more details about inter-annotator agreement, we refer the reader to [9].

[10] http://twitter.com/

[11] http://idebate.org/

[12] http://www.facebook.com/

which are the relations among these arguments, and second they have to understand, given these relations, which are the accepted arguments.

In this section, we answer the following research question: *how to support the participants in natural language (NL) debates to detect which are the relations among the arguments, and which arguments are accepted?* Two kinds of relations connect the arguments in such online debate platforms: a positive relation (i.e., a *support* relation), and a negative relation (i.e., an *attack* relation). To answer to our research question we need to rely on an argumentative framework able to deal with such *bipolar* relations. [133]'s abstract theory defines an argumentation framework as a set of abstract arguments interacting with each others through a so called *attack* relation. In the last years, several proposals to extend the original abstract theory with a *support* relation have been addressed, leading to the birth of *bipolar argumentation* frameworks (BAF) [94], and the further introduction of a number of *additional attacks* among the arguments [93, 55, 237].

Our research question breaks down into the following subquestions:

1. How to automatically identify the arguments, as well as their relationships, from natural language debates?

2. What is the relation between the notion of support in bipolar argumentation and the notion of textual entailment in natural language processing?

First, we propose to combine natural language techniques and Dung-like abstract argumentation to identify and generate the arguments from natural language text, and then to evaluate this set of arguments to know which are the accepted ones. Starting from the participants' opinions, we detect which ones imply or contradict, even indirectly, the issue of the debate using the textual entailment approach. Beside formal approaches to semantic inference that rely on logical representation of meaning, the notion of Textual Entailment (TE) has been proposed as an applied framework to capture major semantic inference needs across applications in the Computational Linguistics field [118]. The development of the Web has witnessed a paradigm shift, due to the need to process a huge amount of available (but often noisy) data. TE is a generic framework for applied semantics, where linguistic objects are mapped by means of semantic inferences at a textual level. We use TE to automatically identify, from a natural language text, the arguments. Second, we adopt bipolar argumentation [94] to reason over the set of generated arguments with the aim of deciding which are the accepted ones. Proposals like argumentation schemes [318], Araucaria [271], Carneades [156], and ArguMed [304] use natural language arguments, but they ask the participants to indicate the semantic relationship among the arguments, and the linguistic content remains unanalyzed. As underlined by [270], "the goal machinery that leads to arguments being automatically generated has been only briefly touched upon, and yet is clearly fundamental to the endeavor". Summarizing, we combine the two approaches, i.e., textual entailment and abstract bipolar argumentation, in a framework whose aim is to *(i)* generate the abstract arguments from the online debates through TE, *(ii)* build the argumentation framework from the arguments and the relationships returned by the TE module, and *(iii)* return the set of accepted arguments. We evaluate the feasibility of our combined approach on a data set extracted from a sample of Debatepedia debates.

Second, we study the relation among the notion of support in bipolar argumentation [94], and the notion of TE in Natural Language Processing (NLP) [118]. In the first study of the current work, we assume the TE relation extracted from NL texts as equivalent to a support relation in bipolar argumentation. This is a strong assumption, and in this second part of our work we aim at verifying on a sample of real data from Debatepedia whether it is always the case that support is equivalent to TE. In particular, for addressing this issue we focus both on the relation between support and entailment, and on the relation between attack and

contradiction. We show that TE and contradiction are more specific concepts than support and attack, but still hold in most of the argument pairs. Moreover, starting from the comparative study addressed by [92], we consider four additional attacks proposed in the literature: *supported* (if argument *a* supports argument *b* and *b* attacks argument *c*, then *a* attacks *c*) and *secondary* (if *a* supports *b* and *c* attacks *a*, then *c* attacks *b*) attacks [93], *mediated* attacks [55] (if *a* supports *b* and *c* attacks *b*, then *c* attacks *a*), and *extended* attacks [238, 237] (if *a* supports *b* and *a* attacks *c*, then *b* attacks *c*). We investigate the presence and the distribution of these attacks in NL debates on a data set extracted from Debatepedia, and we show that all these models are verified in human debates, even if with a different frequency.

The originality of the proposed framework consists in the combination of two techniques which need each other to provide a complete reasoning model: TE has the power to automatically identify the arguments in the text and to specify which kind of relation links each couple of arguments, but it cannot assess which are the *winning* arguments. This is addressed by argumentation theory which lacks automatic techniques to extract the arguments from free text. The combination of these two approaches leads to the definition of a powerful tool to reason over online debates. In addition, the benefit of the proposed deeper analysis of the relation among the two notions of support and TE is twofold. First, it is used to verify, through a data driven evaluation, the "goodness" of the proposed models of bipolar argumentation to be used in real settings, going beyond *ad hoc* NL examples. Second, it can be used to guide the construction of cognitive agents whose major need is to achieve a behavior as close as possible to the human one.

NLP approaches to semantic inference

Classical approaches to semantic inference rely on logical representations of meaning that are external to the language itself, and are typically independent of the structure of any particular natural language. Texts are first translated, or interpreted, into some logical form and then new propositions are inferred from interpreted texts by a logical theorem prover. But, especially after the development of the Web, we have witnessed a paradigm shift, due to the need to process a huge amount of available (but often noisy) data. Addressing the inference task by means of logical theorem provers in automated applications aimed at natural language understanding has shown several intrinsic limitations [48]. As highlighted in [232], in formal approaches semanticists generally opt for rich (i.e. including at least first order logic) representation formalisms to capture as many relevant aspects of the meaning as possible, but practicable methods for generating such representations are very rare. The translation of real-world sentences into logic is difficult because of issues such as ambiguity or vagueness [259]. Moreover, the computational costs of deploying first-order logic theorem prover tools in real world situations may be prohibitive, and huge amounts of additional linguistic and background knowledge are required. Formal approaches address forms of deductive reasoning, and therefore often exhibit a too high level of precision and strictness as compared to human judgments, that allow for uncertainties typical of inductive reasoning [58]. While it is possible to model elementary inferences on the precise level allowed by deductive systems, many pragmatic aspects that play a role in everyday inference cannot be accounted for. Inferences that are plausible but not logically stringent cannot be modeled in a straightforward way, but in NLP applications approximate reasoning should be preferred in some cases to having no answers at all.

Especially in data-driven approaches, like the one sought in this work, where patterns are learnt from large-scale naturally-occurring data, we can settle for approximate answers provided by efficient and robust systems, even at the price of logic unsoundness or incompleteness. Starting from these considerations, [232] propose to address the inference task directly at the textual level instead, exploiting currently available NLP techniques. While methods for automated deduction assume that the arguments in input are already

expressed in some formal meaning representation (e.g. first order logic), addressing the inference task at a textual level opens different and new challenges from those encountered in formal deduction. Indeed, more emphasis is put on informal reasoning, lexical semantic knowledge, and variability of linguistic expressions.

The notion of Textual Entailment has been proposed as an applied framework to capture major semantic inference needs across applications in NLP [118]. It is defined as a relation between a coherent textual fragment (the Text T) and a language expression, which is considered as the Hypothesis (H). Entailment holds (i.e. $T \Rightarrow H$) if the meaning of H can be inferred from the meaning of T, as interpreted by a typical language user. The TE relationship is directional, since the meaning of one expression may usually entail the other, while the opposite is much less certain. Consider the pairs in Example 31 and 32.

Example 31.

T1: *Internet access is essential now; must be a right. The internet is only that wire that delivers freedom of speech, freedom of assembly, and freedom of the press in a single connection.*
H: *Making Internet a right only benefits society.*

Example 32 (Continued)**.**

T2: *Internet not as important as real rights. We may think of such trivial things as a fundamental right, but consider the truly impoverished and what is most important to them. The right to vote, the right to liberty and freedom from slavery or the right to elementary education.*
H: *Making Internet a right only benefits society.*

A system aimed at recognizing TE should detect an inference relation between T1 and H (i.e. the meaning of H can be derived from the meaning of T) in Example 31, while it should not detect an entailment between T2 and H in Example 32. As introduced before, TE definition is based on (and assumes) common human understanding of language, as well as common background knowledge. However, the entailment relation is said to hold only if the statement in the text licenses the statement in the hypothesis, meaning that the content of T and common knowledge together should entail H, and not background knowledge alone. In this applied framework, inferences are performed directly over lexical-syntactic representations of the texts. Such definition of TE captures quite broadly the reasoning about language variability needed by different applications aimed at NL understanding and processing, e.g. information extraction [275] and text summarization [24]. Differently from the classical semantic definition of entailment [106], the notion of TE accounts for some degree of uncertainty allowed in applications (see Example 31).

In 2005, the PASCAL Network of Excellence started an attempt to promote a generic evaluation framework covering semantic-oriented inferences needed for practical applications, launching the Recognizing Textual Entailment challenge [118], with the aim of setting a unifying benchmark for the development and evaluation of methods that typically address similar problems in different, application-oriented, manners. As many of the needs of several NLP applications can be cast in terms of TE, the goal of the evaluation campaign is to promote the development of general entailment recognition engines, designed to provide generic modules across applications. Since 2005, such initiative has been repeated yearly[13], asking the participants to develop a system that, given two text fragments (the *text* T and the *hypothesis* H), can determine whether the meaning of one text is entailed, i.e. can be inferred, from the other. For pairs where

[13] `http://aclweb.org/aclwiki/index.php?title=Recognizing_Textual_Entailment`

the entailment relation does not hold between T and H, systems are required to make a further distinction between pairs where the entailment does not hold because the content of H is contradicted by the content of T (i.e. *contradiction*, see Example 32), and pairs where the entailment cannot be determined because the truth of H cannot be verified on the basis of the content of T (i.e. *unknown*, see Example 33). [215] provide a definition of contradiction for the TE task, claiming that it occurs when two sentences *i)* are extremely unlikely to be true simultaneously, and *ii)* involve the same event. This three-way judgment task (*entailment* vs *contradiction* vs *unknown*) was introduced since RTE-4, while before a two-way decision task (*entailment* vs *no entailment*) was asked to participating systems. However, the classic two-way task is offered as an alternative also in recent editions of the evaluation campaign (*contradiction* and *unknown* judgments are collapsed into the judgment *no entailment*).

In our work, we consider the three way scenario to map TE relation with bipolar argumentation, focusing both on the relation between support and entailment, and on the relation between attack and contradiction. As will be discussed later on in this section, we consider argument pairs connected by a relation of support (but where the first argument does not entail the second one), and argument pairs connected by a relation of attack (but where the first argument does not contradict the second one) as *unknown* pairs in the TE framework.

Example 33 (Continued)**.**

T3: *Internet "right" means denying parents' ability to set limits. Do you want to make a world when a mother tells her child: "you cannot stay on the internet anymore" that she has taken a right from him? Compare taking the right for a home or for education with taking the "right" to access the internet.*

H: *Internet access is essential now; must be a right. The internet is only that wire that delivers freedom of speech, freedom of assembly, and freedom of the press in a single connection.*

The systems submitted to the RTE challenge are tested against manually annotated data sets, which include typical examples that correspond to success and failure cases of NLP applications. A number of data-driven approaches applied to semantics have been experimented throughout the years. In general, the approaches still more used by the submitted systems include Machine Learning (typically SVM), logical inference, cross-pair similarity measures between T and H, and word alignment - for an overview, see [5], and [118].

Bipolar argumentation

This section provides the basic concepts of bipolar argumentation [94].

Bipolar argumentation frameworks, firstly proposed by [94], extend Dung's framework taking into account both the attack relation and the support relation. In particular, an abstract bipolar argumentation framework is a labeled directed graph, with two labels indicating either attack or support. In this section, we represent the attack relation by $a \rightarrow b$, and the support relation by $a \Rightarrow b$.

Definition 63. (Bipolar argumentation framework) A bipolar argumentation framework (BAF) is a tuple $\langle Ar, \rightarrow, \Rightarrow \rangle$ where A is the set of elements called arguments, and two binary relations over Ar are called *attack* and *support*, respectively.

[92] address a formal analysis of the models of support in bipolar argumentation to achieve a better understanding of this notion and its uses. [94, 93] argue about the emergence of new kinds of attacks from the

interaction between attacks and supports in BAF. In the rest of the section, we will adopt their terminology to refer to additional attacks, i.e., *complex attacks*. In particular, they specify two kinds of complex attacks called *secondary* and *supported* attacks, respectively.

Definition 64. (Secondary and supported attacks) Let BAF = $\langle Ar, \rightarrow, \Rightarrow \rangle$ where $a, b \in Ar$. A *supported* attack for b by a is a sequence $a_1 R_1 \ldots R_{n-1} a_n$, $n \geq 3$, with $a_1 = a, a_n = b$, such that $\forall i = 1 \ldots n-2, R_i = \Rightarrow$ and $R_{n-1} = \rightarrow$. A secondary attack for b by a is a sequence $a_1 R_1 \ldots R_{n-1} a_n$, $n \geq 3$, with $a_1 = a, a_n = b$, such that $R_1 = \rightarrow$ and $\forall i = 2 \ldots$ n-1,$R_i = \Rightarrow$.

According to the above definition, these attacks hold in the first two cases depicted in Figure 5.1, where there is a supported attack from a to c, and there is a secondary attack from c to b. In this section, we represent complex attacks using a dotted arrow.

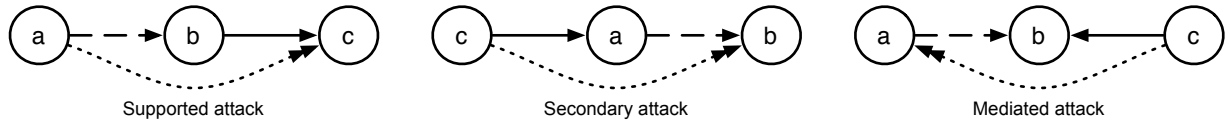


Figure 5.1: Additional attacks emerging from the interaction of supports and attacks.

The support relation has been specialized in other approaches where new complex attacks emerging from the combination of existing attacks and supports are proposed. [55] propose a *deductive* view of support in abstract argumentation where, given the support $a \Rightarrow b$ the acceptance of a implies the acceptance of b, and the rejection of b implies the rejection of a. They introduce a new kind of complex attacks called *mediated* attacks (Figure 5.1).

Definition 65. (Mediated attacks) Let BAF = $\langle Ar, \rightarrow, \Rightarrow \rangle$ where $a, b \in Ar$. A mediated attack on b by a is a sequence $a_1 R_1 \ldots R_{n-2} a_{n-1}$ and $a_n R_{n-1} a_{n-1}$, $n \geq 3$, with $a_1 = a, a_{n-1} = b, a_n = c$, such that $R_{n-1} = \rightarrow$ and $\forall i = 1 \ldots$ n-2,$R_i = \Rightarrow$.

[238, 237] propose, instead, an account of support called *necessary* support. In this framework, given $a \Rightarrow b$ then the acceptance of a is necessary to get the acceptance of b, i.e., the acceptance of b implies the acceptance of a. They introduce two new kinds of complex attacks called *extended attacks* (Figure 5.1). Note that the first kind of extended attacks is equivalent to the secondary attacks introduced by [94, 93], and that the second case is the dual of supported attacks. See [92] for a formal comparison of the different models of support in bipolar argumentation.

Definition 66. (Extended attacks) Let BAF = $\langle Ar, \rightarrow, \Rightarrow \rangle$ where $a, b \in Ar$. An extended attack on b by a is a sequence $a_1 R_1 a_2 R_2 \ldots R_n a_n$, $n \geq 3$, with $a_1 = a, a_n = b$, such that $R_1 = \rightarrow$ and $\forall i = 2 \ldots$ n,$R_i = \Rightarrow$, or a sequence $a_1 R_1 \ldots R_n a_n$ and $a_1 R_p a_p$, $n \geq 2$, with $a_n = a, a_p = b$, such that $R_p = \rightarrow$ and $\forall i = 1 \ldots$ n,$R_i = \Rightarrow$.

All these models of support in bipolar argumentation address the problem of how to compute the set of extensions from the extended framework providing different kinds of solutions, i.e., introducing the notion of *safety* in BAF [94], or computing the extensions in the meta-level [55, 93]. In this section, we are not interested in discussing and evaluating these different solutions. Our aim is to evaluate how much these different models of support occur and are effectively "exploited" in natural language dialogues, towards a better understanding of the notion of support and attack in bipolar argumentation.

We are aware that the notion of support is controversial in the field of argumentation theory. In particular, another view of support sees this relation as a relation holding among the premises and the conclusion of a structured argument, and not as another relation among atomic arguments [263]. However, given the amount of attention bipolar argumentation is receiving in the literature [266], a better account of this kind of frameworks is required.

Another approach to model support has been proposed by [241] and [242], where they distinguish among *prima-facie* arguments and standard ones. They show how a set of arguments described using Dung's argumentation framework can be mapped from and to an argumentation framework that includes both attack and support relations. The idea is that an argument can be accepted only if there is an evidence supporting it, i.e., evidence is represented by means of *prima-facie* arguments. In this section, we do not intend to take a position in this debate. We focus our analysis on the abstract models of bipolar argumentation proposed in the literature [93, 55, 237], and we leave as future work the account of support in structured argumentation and the model proposed by [241] and [242].

Casting bipolar argumentation as a TE problem

The goal of our work is to propose an approach to support the participants in forums or debates (e.g. Debatepedia, Twitter) to detect which arguments among the ones expressed by the other participants on a certain topic are accepted. As a first step, we need to *(i)* automatically generate the arguments (i.e. recognize a participant's opinion on a certain topic as an argument), as well as *(ii)* detect their relation with respect to the other arguments. We cast the described problem as a TE problem, where the T-H pair is a pair of arguments expressed by two different participants in a debate on a certain topic. For instance, given the argument "Making Internet a right only benefits society" (that we consider as H as a starting point), participants can be in favor of it (expressing arguments from which H can be inferred, as in Example 31), or can contradict such argument (expressing an opinion against it, as in Example 32). Since in debates one participant's argument comes after the other, we can extract such arguments and compare them both w.r.t. the main issue, and w.r.t. the other participants' arguments (when the new argument entails or contradicts one of the arguments previously expressed by another participant). For instance, given the same debate as before, a new argument T3 may be expressed by a third participant to contradict T2 (that becomes the new H (H1) in the pair), as shown in Example 34.

Example 34 (Continued)**.**

T3: *I've seen the growing awareness within the developing world that computers and connectivity matter and can be useful. It's not that computers matter more than water, food, shelter and healthcare, but that the network and PCs can be used to ensure that those other things are available. Satellite imagery sent to a local computer can help villages find fresh water, mobile phones can tell farmers the prices at market so they know when to harvest.*

T2 $\equiv$ **H1**: *Internet not as important as real rights. We may think of such trivial things as a fundamental right, but consider the truly impoverished and what is most important to them. The right to vote, the right to liberty and freedom from slavery or the right to elementary education.*

With respect to the goal of our work, TE provides us with the techniques to identify the arguments in a debate, and to detect which kind of relation underlies each couple of arguments. A TE system returns indeed a judgment (entailment or contradiction) on the arguments pairs related to a certain topic, that are used as

input to build the argumentation framework, as described in the next section. Example 35 presents how we combine TE with bipolar argumentation to compute at the end the set of accepted arguments.

Example 35 (Continued)**.**

The textual entailment phase returns the following couples for the natural language opinions detailed in Examples 31, 32, and 34:

- T1 entails H
- T2 attacks H
- T3 attacks H1 (i.e., T2)

Given this result, the argumentation module of our framework maps each element to its corresponding argument: H $\equiv A_1$, T1 $\equiv A_2$, T2 $\equiv A_3$, and T3 $\equiv A_4$. The resulting argumentation framework, visualized in Figure 5.2, shows that the accepted arguments (using admissibility-based semantics) are $\{A_1, A_2, A_4\}$. This means that the issue "Making Internet a right only benefits society"' A_1 is considered as accepted. Double bordered arguments are the accepted ones.

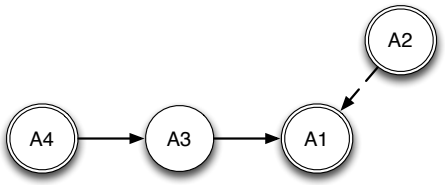


Figure 5.2: The argumentation framework built from the results of the TE module for Examples 31, 32, and 34.

Experimental setting

As a case study to experiment the combination of TE and argumentation theory to support the interaction of participants in online debates, we select Debatepedia, an encyclopedia of pro and con arguments on critical issues. First, we describe the creation of the data set of T-H pairs extracted from a sample of Debatepedia topics, then we present the TE system we use, and we report on obtained results.

Data set. To create the data set of arguments pairs to evaluate our task, we follow the criteria defined and used by the organizers of RTE (see Section 5.3). To test the progress of TE systems in a comparable setting, the participants to RTE are provided with data sets composed of T-H pairs involving various levels of entailment reasoning (e.g. lexical, syntactic), and TE systems are required to produce a correct judgment on the given pairs (i.e. to say if the meaning of one text snippet can be inferred from the other). The data available for the RTE challenges are not suitable for our goal, since the pairs are extracted from news and are not linked among each others (i.e. they do not report opinions on a certain topic).

For this reason, we created a data set to evaluate our combined approach focusing on Debatepedia. We manually selected a set of topics (Table 5.3 column *Topics*) of Debatepedia debates, and for each topic we apply the following procedure:

1. the main issue (i.e., the title of the debate in its affirmative form) is considered as the starting argument;
2. each user opinion is extracted and considered as an argument;
3. since *attack* and *support* are binary relations, the arguments are coupled with:
   a) the starting argument, or
   b) other arguments in the same discussion to which the most recent argument refers (i.e., when a user opinion supports or attacks an argument previously expressed by another user, we couple the former with the latter), following the chronological order to maintain the dialogue structure;
4. the resulting pairs of arguments are then tagged with the appropriate relation, i.e., *attack* or *support*[14].

Using Debatepedia as case study provides us with already annotated arguments (*pro* $\Rightarrow$ *entailment*[15], and *con* $\Rightarrow$ *contradiction*), and casts our task as a yes/no entailment task. To show a step-by-step application of the procedure, let us consider the debated issue *Can coca be classified as a narcotic?*. At step 1, we transform its title into the affirmative form, and we consider it as the starting argument (a). Then, at step 2, we extract all the users opinions concerning this issue (both pro and con), e.g., (b), (c) and (d):

Example 36.
(a) *Coca can be classified as a narcotic.*

(b) *In 1992 the World Health Organization's Expert Committee on Drug Dependence (ECDD) undertook a "prereview" of coca leaf at its 28th meeting. The 28th ECDD report concluded that, "the coca leaf is appropriately scheduled as a narcotic under the Single Convention on Narcotic Drugs, 1961, since cocaine is readily extractable from the leaf." This ease of extraction makes coca and cocaine inextricably linked. Therefore, because cocaine is defined as a narcotic, coca must also be defined in this way.*

(c) *Coca in its natural state is not a narcotic. What is absurd about the 1961 convention is that it considers the coca leaf in its natural, unaltered state to be a narcotic. The paste or the concentrate that is extracted from the coca leaf, commonly known as cocaine, is indeed a narcotic, but the plant itself is not.*

(d) *Coca is not cocaine. Coca is distinct from cocaine. Coca is a natural leaf with very mild effects when chewed. Cocaine is a highly processed and concentrated drug using derivatives from coca, and therefore should not be considered as a narcotic.*

At step 3a we couple the arguments (b) and (d) with the starting issue since they are directly linked with it, and at step 3b we couple argument (c) with argument (b), and argument (d) with argument (c) since they follow one another in the discussion (i.e. user expressing argument (c) answers back to user expressing argument (b), so the arguments are concatenated - the same for arguments (d) and (c)).
At step 4, the resulting pairs of arguments are then tagged with the appropriate relation: **(b)** *supports* **(a)**, **(d)** *attacks* **(a)**, **(c)** *attacks* **(b)** and **(d)** *supports* **(c)**.

We collected 200 T-H pairs (Table 5.3), 100 to train and 100 to test the TE system (each data set is composed by 55 entailment and 45 contradiction pairs). The pairs considered for the test set concern completely new topics, never seen by the system.

[14]The data set is freely available at `http://bit.ly/debatepedia_ds`.

[15]Here we consider only arguments implying another argument. Arguments "supporting" another argument, but not inferring it will be discussed in Section 5.3.

Training set				
Topic	**#argum**	**#pairs**		
		TOT.	**yes**	**no**
Violent games boost aggressiveness	16	15	8	7
China one-child policy	11	10	6	4
Consider coca as a narcotic	15	14	7	7
Child beauty contests	12	11	7	4
Arming Libyan rebels	10	9	4	5
Random alcohol breath tests	8	7	4	3
Osama death photo	11	10	5	5
Privatizing social security	11	10	5	5
Internet access as a right	15	14	9	5
TOTAL	109	**100**	**55**	**45**
Test set				
Topic	**#argum**	**#pairs**		
		TOT.	**yes**	**no**
Ground zero mosque	9	8	3	5
Mandatory military service	11	10	3	7
No fly zone over Libya	11	10	6	4
Airport security profiling	9	8	4	4
Solar energy	16	15	11	4
Natural gas vehicles	12	11	5	6
Use of cell phones while driving	11	10	5	5
Marijuana legalization	17	16	10	6
Gay marriage as a right	7	6	4	2
Vegetarianism	7	6	4	2
TOTAL	110	**100**	**55**	**45**

Table 5.1: The Debatepedia data set used in our experiments.

TE system. To detect which kind of relation underlies each couple of arguments, we take advantage of the modular architecture of the EDITS system (Edit Distance Textual Entailment Suite) version 3.0, an open-source software package for recognizing TE[16] [189]. EDITS implements a distance-based framework which assumes that the probability of an entailment relation between a given T-H pair is inversely proportional to the distance between T and H (i.e., the higher the distance, the lower is the probability of entailment).[17] Within this framework the system implements different approaches to distance computation, i.e., both edit distance algorithms (that calculate the T-H distance as the cost of the edit operations, i.e., insertion, deletion

[16] `http://edits.fbk.eu/`

[17] In previous RTE challenges, EDITS always ranked among the 5 best participating systems out of an average of 25 systems, and is one of the few RTE systems available as open source `http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool`

		Train			Test		
	rel	*Pr.*	*Rec.*	*Acc.*	*Pr.*	*Rec.*	*Acc.*
EDITS	*yes*	0.71	0.73	**0.69**	0.69	0.72	**0.67**
	no	0.66	0.64		0.64	0.6	
WordOverl.	*yes*	0.64	0.65	0.61	0.64	0.67	0.62
	no	0.56	0.55		0.58	0.55	

Table 5.2: Systems performances on the Debatepedia data set (precision, recall and accuracy)

and substitution that are necessary to transform T into H), and similarity algorithms. Each algorithm returns a normalized distance score. At a training stage, distance scores calculated over annotated T-H pairs are used to estimate a threshold that best separates positive from negative examples. Such threshold is then used at a test stage to assign a judgment and a confidence score to each test pair.

Evaluation. To evaluate our combined approach, we carry out a two-step evaluation: first, we assess the performances of the TE system to correctly assign the entailment and contradiction relations to the pairs of arguments in the Debatepedia data set. Then, we evaluate how much such performances impact on the application of the argumentation theory module, i.e. how much a wrong assignment of a relation to a pair of arguments is propagated in the argumentation framework.

For the first evaluation, we run EDITS on the Debatepedia training set to learn the model, and we test it on the test set. We tuned EDITS in the following configuration: *i)* cosine similarity as the core distance algorithm, *ii)* distance calculated on lemmas, and *iii)* a stopword list is defined to set no distance between stopwords. We use the system off-the-shelf, applying one of its basic configurations. As future work, we plan to fully exploit EDITS features, integrating background and linguistic knowledge in the form of entailment rules, and to calculate the distance between T and H on their syntactic structure.

Table 5.2 reports on the obtained results both using EDITS and using a baseline that applies a Word Overlap algorithm on tokenized text. Even using a basic configuration of EDITS, and a small data set (100 pairs for training) performances on Debatepedia test set are promising, and in line with performances of TE systems on RTE data sets (usually containing about 1000 pairs for training and 1000 for test). In order to understand if increasing the number of argument pairs in the training set could bring to an improvement in the system performances, the EDITS learning curve is visualized in Figure 5.3. Note that augmenting the number of training pairs actually improves EDITS accuracy on the test set, meaning that we should consider extending the Debatepedia data set for future work.

As a second step in our evaluation phase, we consider the impact of EDITS performances on the acceptability of the arguments, i.e. how much a wrong assignment of a relation to a pair of arguments affects the acceptability of the arguments in the argumentation framework. We use admissibility-based semantics to identify the accepted arguments both on the correct argumentation framework of each Debatepedia topic (where entailment/contradiction relations are correctly assigned, i.e. the goldstandard), and on the framework generated assigning the relations resulted from the TE system judgments. The precision of the combined approach we propose in the identification of the accepted arguments is on average 0.74 (i.e. arguments accepted by the combined system and by the goldstandard w.r.t. a certain Debatepedia topic), and the recall is 0.76 (i.e. arguments accepted in the goldstandard and retrieved as accepted by the combined

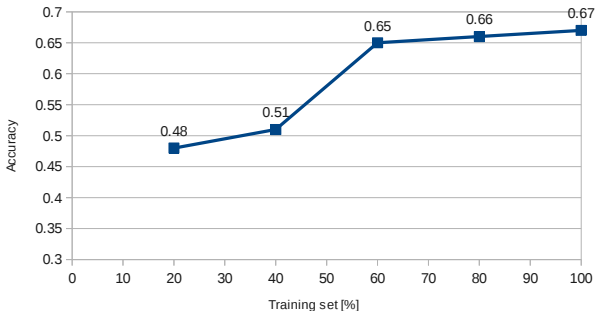


Figure 5.3: EDITS learning curve on Debatepedia data set

system). Its accuracy (i.e. ability of the combined system to accept some arguments and discard some others) is 0.75, meaning that the TE system mistakes in relation assignment propagate in the argumentation framework, but results are still satisfying.

Extending the analysis on bipolar argumentation beyond TE

In the previous section, we assumed the TE relation extracted from NL texts as equivalent to the support relation in bipolar argumentation. On closer view, this is a strong assumption. In this second part of our work, we aim at verifying on an extended sample of real data from Debatepedia whether it is always the case that support is equivalent to TE. In particular, for addressing this issue, we focus both on the relations between support and entailment, and on the relations between attack and contradiction. We extend the data set we presented, extracting an additional set of arguments from Debatepedia topics. Even if our data set cannot be exhaustive, the methodology we apply for the arguments extraction aims at preserving the original structure of the debate, to make it as representative as possible of daily human interactions in natural language.

Two different empirical studies are presented in this section. The first one follows the analysis presented in Section 5.3, and explores the relation among the notion of *support* and *attack* in bipolar argumentation, and the *semantic inferences* as defined in NLP. The second analysis starts instead from the comparative study of [92] of the four complex attacks proposed in the literature (see Section 5.3), and investigates their distribution in NL debates.

Data set. We select the same topics as for the first version of the dataset, since this is the only freely available data set of natural language arguments (Table 5.3, column *Topics*). But Since that data set was created respecting the assumption that the TE relation and the support relation are equivalent, in all the previously collected pairs both TE and support relations (or contradiction and attack relations) hold.

In this study we want to move a step further, to understand whether it is always the case that support is equivalent to TE (and contradiction to attack). We therefore apply again the extraction methodology described before to extend our data set. In total, our new data set contains 310 different arguments and 320 argument pairs (179 expressing the *support* relation among the involved arguments, and 141 expressing the *attack* relation, see Table 5.3). We consider the obtained data set as representative of human debates in a non-controlled setting (Debatepedia users position their arguments with respect to the others as PRO or CON, the data are not biased), and we use it for our empirical studies.

DEBATEPEDIA data set		
Topic	**#argum**	**#pairs**
VIOLENT GAMES BOOST AGGRESSIVENESS	17	23
CHINA ONE-CHILD POLICY	11	14
CONSIDER COCA AS A NARCOTIC	17	22
CHILD BEAUTY CONTESTS	13	17
ARMING LIBYAN REBELS	13	15
RANDOM ALCOHOL BREATH TESTS	11	14
OSAMA DEATH PHOTO	22	24
PRIVATIZING SOCIAL SECURITY	12	13
INTERNET ACCESS AS A RIGHT	15	17
GROUND ZERO MOSQUE	11	12
MANDATORY MILITARY SERVICE	15	17
NO FLY ZONE OVER LIBYA	18	19
AIRPORT SECURITY PROFILING	12	13
SOLAR ENERGY	18	19
NATURAL GAS VEHICLES	16	17
USE OF CELL PHONES WHILE DRIVING	16	16
MARIJUANA LEGALIZATION	23	25
GAY MARRIAGE AS A RIGHT	10	10
VEGETARIANISM	14	13
TOTAL	310	**320**

Table 5.3: Debatepedia data set.

First study: support and TE. Our first empirical study aims at a better understanding of the relation among the notion of support in bipolar argumentation [92], and the definition of semantic inference in NLP (in particular, the more specific notion of TE) [118].

Basing on the TE definition, an annotator with skills in linguistics has carried out a first phase of annotation of the Debatepedia data set. The goal of such annotation is to individually consider each pair of *support* and *attack* among arguments, and to additionally tag them as *entailment*, *contradiction* or *null*. The *null* judgment can be assigned in case an argument is supporting another argument without inferring it, or the argument is attacking another argument without contradicting it. As exemplified in Example 36, a correct entailment pair is **(b)** $\Rightarrow$ **(a)**, while a contradiction is **(d)** $\not\Rightarrow$ **(a)**. A *null* judgment is assigned to **(d) - (c)**, since the former argument supports the latter without inferring it. Our data set is an extended version of [72]'s one allowing for a deeper investigation.

To assess the validity of the annotation task, we calculate the inter-annotator agreement. Another annotator with skills in linguistics has therefore independently annotated a sample of 100 pairs of the data set. We calculated the inter-annotator agreement considering the argument pairs tagged as *support* and *attacks* by both annotators, and we verify the agreement between the pairs tagged as *entailment* and as *null* (i.e. no entailment), and as *contradiction* and as *null* (i.e. no contradiction), respectively. Applying κ to our data, the agreement for our task is $\kappa = 0.74$. As a rule of thumb, this is a satisfactory agreement.

Table 5.4 reports the results of the annotation on our Debatepedia data set, as resulting after a reconcili-

ation phase carried out by the annotators[18].

Relations		% arguments (# arg.)
support	+ *entailment*	61.6 (111)
	- *entailment (null)*	38.4 (69)
attack	+ *contradiction*	71.4 (100)
	- *contradiction (null)*	28.6 (40)

Table 5.4: Support and TE relations on Debatepedia data set.

On the 320 pairs of the data set, 180 represent a *support* relation, while 140 are *attacks*. Considering only the *supports*, we can see that 111 argument pairs (i.e., 61.6%) are an actual entailment, while in 38.4% of the cases the first argument of the pair supports the second one without inferring it (e.g. **(d) - (c)** in Example 36). With respect to the *attacks*, we can notice that 100 argument pairs (i.e., 71.4%) are both attack and contradiction, while only the 28.6% of the argument pairs does not contradict the arguments they are attacking, as in Example 37.

Example 37.
(e) *Coca chewing is bad for human health. The decision to ban coca chewing fifty years ago was based on a 1950 report elaborated by the UN Commission of Inquiry on the Coca Leaf with a mandate from ECOSOC: "We believe that the daily, inveterate use of coca leaves by chewing is thoroughly noxious and therefore detrimental".*
(f) *Chewing coca offers an energy boost. Coca provides an energy boost for working or for combating fatigue and cold.*

Differently from the relation between support-entailment, the difference between attack and contradiction is more subtle, and it is not always straightforward to say whether an argument attacks another argument without contradicting it. In Example 37, we consider that **(e)** does not contradict **(f)** even if it attacks **(f)**, since chewing coca can offer an energy boost, and still be bad for human health. This kind of attacks is less frequent than the attacks-contradictions (see Table 5.4).

Considering the three way scenario to map TE relation with bipolar argumentation, argument pairs connected by a relation of support (but where the first argument does not entail the second one), and argument pairs connected by a relation of attack (but where the first argument does not contradict the second one) have to be mapped as *unknown* pairs in the TE framework. The *unknown* relation in TE refers to the T-H pairs where the entailment cannot be determined because the truth of H cannot be verified on the basis of the content of T. This is a broad definition, that can apply also to pairs of non related sentences (that are considered as unrelated arguments in bipolar argumentation).

From an application viewpoint, as highlighted in [270] and [167], argumentation theory should be used as a tool in on-line discussions applications to identify the relations among the statements, and provide a structure to the dialogue to easily evaluate the user's opinions. Starting from the methodology proposed in Section 5.3 for passing from natural language arguments to a bipolar argumentation framework, our study

[18]In this phase, the annotators discuss the results to find an agreement on the annotation to be released.

demonstrates that applying the TE approach would be productive in the 66% of the Debatepedia data set. Other techniques should then be experimented to cover the other cases, for instance measuring the semantic relatedness of the two propositions using Latent Semantics Analysis techniques [191].

Second study: complex attacks. We carry out now a comparative evaluation of the four additional attacks proposed in the literature, and we investigate their meaning and distribution on the sample of NL arguments.

Basing on the additional attacks (Section 5.3), and the original AF of each topic in our data set (Table 5.3), the following procedure is applied: the *supported* (secondary, mediated, and extended, respectively) attacks are added, and the argument pairs resulting from coupling the arguments linked by this relation are collected in the data set "supported (secondary, mediated, and extended, respectively) attack". Collecting the argument pairs generated from the different types of complex attacks in separate data sets allows us to independently analyze each type, and to perform a more accurate evaluation.[19] Figures 5.4a-d show the four AFs resulting from the addition of the complex attacks in the example *Can coca be classified as a narcotic?*. Note that the AF in Figure 5.4a, where the supported attack is introduced, is the same of Figure 5.4b where the mediated attack is introduced. Notice that, even if the additional attack which is introduced coincide, i.e., *d* attacks *b*, this is due indeed to different interactions among supports and attacks (as highlighted in the figure), i.e., in the case of supported attacks this is due to the support from *d* to *c* and the attack from *c* to *b*, while in the case of mediated attacks this is due to the support from *b* to *a* and the attack from *d* to *a*.

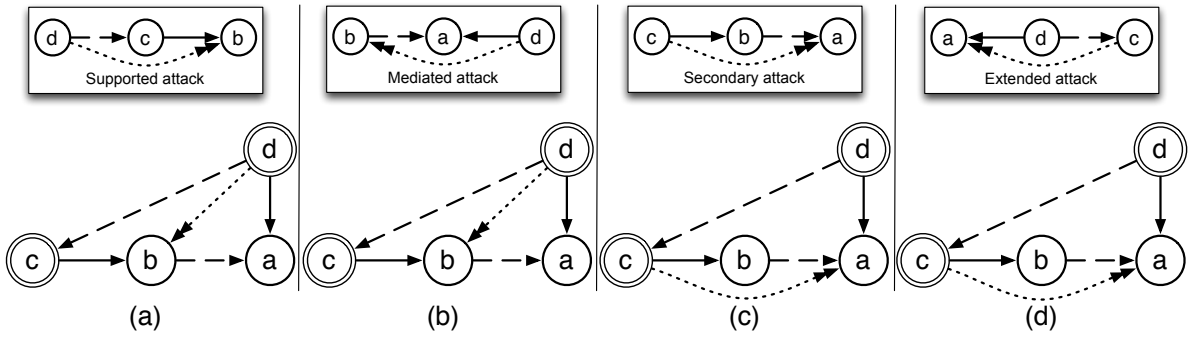


Figure 5.4: The bipolar argumentation framework with the introduction of complex attacks. The top figures show which combination of support and attack generates the new additional attack.

A second annotation phase is then carried out on the data set, to verify if the generated argument pairs of the four data sets are actually attacks (i.e., if the models of complex attacks proposed in the literature are represented in real data). More specifically, an argument pair resulting from the application of a complex attack can be annotated as: *attack* (if it is a correct attack) or as *unrelated* (in case the meanings of the two arguments are not in conflict). For instance, the argument pair **(g)-(h)** (Example 38) resulting from the insertion of a *supported* attack, cannot be considered as an attack since the arguments are considering two different aspects of the issue.

Example 38. (g) *Chewing coca offers an energy boost. Coca provides an energy boost for working or for combating fatigue and cold.*

[19] Data sets freely available for research purposes at `http://bit.ly/VZIs6M`

(h) *Coca can be classified as a narcotic.*

In the annotation, *attacks* are then annotated also as *contradiction* (if the first argument contradicts the other) or *null* (in case the first argument does not contradict the argument it is attacking, as in Example 37). Due to the complexity of the annotation, the same annotation task has been independently carried out also by a second annotator, so as to compute inter-annotator agreement. It has been calculated on a sample of 80 argument pairs (20 pairs randomly extracted from each of the "complex attacks" data set), and it has the goal to assess the validity of the annotation task (counting when the judges agree on the same annotation). We calculated the inter-annotator agreement for our annotation task in two steps. We *(i)* verify the agreement of the two judges on the argument pairs classification *attacks*/*unrelated*, and *(ii)* consider only the argument pairs tagged as *attacks* by both annotators, and we verify the agreement between the pairs tagged as *contradiction* and as *null* (i.e. no contradiction). Applying κ to our data, the agreement for the first step is $\kappa = 0.77$, while for the second step $\kappa = 0.71$. As a rule of thumb, both agreements are satisfactory, although they reflect the higher complexity of the second annotation (*contradiction*/*null*), as pointed out before.

The distribution of complex attacks in the Debatepedia data set, as resulting after a reconciliation phase carried out by the annotators, is shown in Table 5.5. As can be noticed, the *mediated* attack is the most frequent type of attack, generating 335 new argument pairs in the NL sample we considered (i.e. the conditions that allow the application of this kind of complex attacks appear more frequently in real debates). Together with *secondary* attacks, they appear in the AFs of all the debated topics. On the contrary, *extended* attacks are added in 11 out of 19 topics, and *supported* attacks in 17 out of 19 topics. Considering all the topics, on average only 6 pairs generated from the additional attacks were already present in the original data set, meaning that considering also these attacks is a way to hugely enrich our data set of NL debates.

Proposed models	**# occ.**	**attacks**		**unrelated**
		+ contr (null)	*- contr (null)*	
Supported attacks	47	23	17	7
Secondary attacks	53	29	18	6
Mediated attacks	335	84	148	103
Extended attacks	28	15	10	3

Table 5.5: Complex attacks distribution in our data set.

Figure 5.5 graphically represents the complex attacks distribution. Considering the first step of the annotation (i.e. *attacks* vs *unrelated*), the figure shows that the latter case is very infrequent, and that (except for *mediated* attacks) on average only 10% of the argument pairs are tagged as *unrelated*. This observation can be considered as a proof of concept of the four theoretical models of complex attacks we analyzed. Due to the fact that the conditions for the application of the *mediated* attacks are verified more often in the data, it has the drawback of generating more unrelated pairs. Still, the number of successful cases is high enough to consider this kind of attack as representative of human interactions. Considering the second step of the annotation (i.e. *attacks* as *contradiction* or *null*), we can see that results are in line with those reported in our first study (Table 5.4), meaning that also among complex attacks the same distribution is maintained.

The research presented in this section is interdisciplinary. We have integrated in a combined framework an approach from computational linguistics and a technique for non-monotonic reasoning. The aim of this

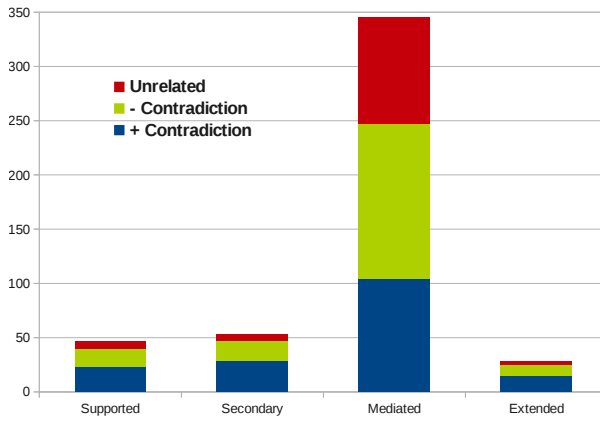


Figure 5.5: Complex attacks distribution in our data set.

research is to provide the participants of online debates and forums with a framework supporting their interaction with the application. In particular, the proposed framework helps the participants to have an overview of the debates, understanding which are the accepted arguments at time being. The key contribution of our research is to allow the automatic detection and generation of the abstract arguments from natural language texts.

5.4 A Support Framework for Argumentative Discussions Management in the Web

On the Social Web, wiki-like platforms allow users to publicly publish their own arguments and opinions. Such arguments are not always accepted by other users on the Web, leading to the publication of additional arguments attacking or supporting the previously proposed ones. The most well known example of such kind of platform is Wikipedia[20] where users may change pieces of text written by other users to support, i.e., further specify them, or attack them, i.e., correcting factual errors or highlighting opposite points of view. Managing such kind of "discussions" using the revision history is a tricky task, and it may be affected by a number of drawbacks. First, the dimension of these discussions makes it difficult for both users and community managers to navigate, and more importantly, understand the meaning of the ongoing discussion. Second, the discussions risk to re-start when newcomers propose arguments which have already been proposed and addressed in the same context. Third, these discussions are not provided in a machine-readable format to be queried by community managers to discover insightful meta-information on the discussions themselves, e.g., discover the number of attacks against arguments about a particular politician concerning the economic growth during his government.

In this section, we answer the following research question: *how to support community managers in managing the discussions on the wiki pages?* This question breaks down into the following subquestions: (i) how to automatically discover the arguments and the relations among them?, and (ii) how to have the overall view of the ongoing discussion to detect the *winning* arguments? The answer to these sub-questions allows us to answer to further questions: how to detect repeated arguments and avoid loops of changes?,

[20] http://en.wikipedia.org/wiki/Main_Page

and how to discover further information on the discussion history? Approaches such as the lightweight vocabulary SIOC Argumentation [193] provide means to model argumentative discussions of social media sites, but they are not able to automatically acquire information about the argumentative structures. As underlined by Lange et al. [193], such a kind of automatic annotation needs the introduction of Natural Language Processing (NLP) techniques to automatically detect the arguments in the texts.

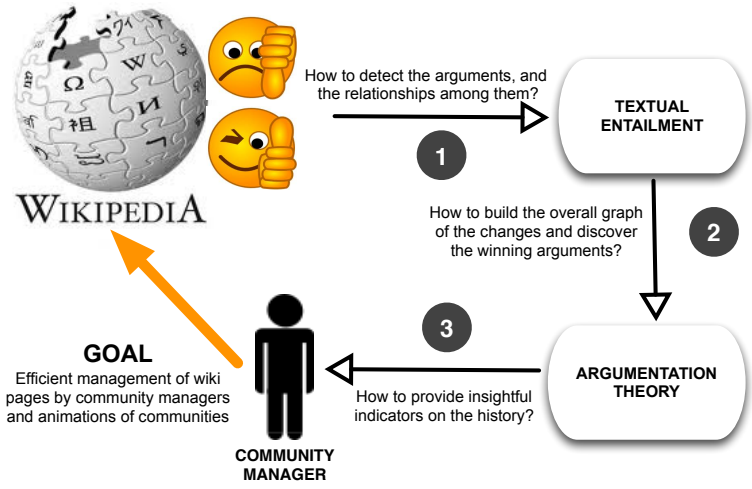


Figure 5.6: An overview of the proposed approach to support community managers.

In this work, we propose a combined framework where a natural language module that automatically detects the arguments and their relations (i.e. *support* or *challenge*), is coupled with an argumentation module to have the overall view of the discussion and detect the winning arguments, as visualized in Figure 5.6.

First, to automatically detect natural language arguments and their relations, we rely on the Textual Entailment (TE) framework, proposed as an applied model to capture major semantic inference needs across applications in the NLP field [118]. Differently from formal approaches to semantic inference, in TE linguistic objects are mapped by means of semantic inferences at a textual level.

Second, we adopt abstract argumentation theory [132] to unify the results of the TE module into a unique argumentation framework able not only to provide the overall view of the discussion, but also to detect the set of *accepted* arguments relying on argumentation semantics. Argumentation theory aims at representing the different opinions of the users in a structured way to support decision making.

Finally, the generated argumentative discussions are described using an extension of the SIOC Argumentation vocabulary[21] thus providing a machine readable version. Such discussions expressed using RDF allow the extraction of a kind of "meta-information" by means of queries, e.g., in SPARQL. These meta-information cannot be easily detected by human users without the support of our automatic framework.

The aim of the proposed framework is twofold: on one side, we want to provide a support to community managers for notification and reporting, e.g., notify the users when their own arguments are attacked, and on the other hand, we support community managers to extract further insightful information from the argumentative discussions. As a case study, we apply and experiment our framework on Wikipedia revision history over a four-year period, focusing in particular on the top five most revised articles.

[21] `http://rdfs.org/sioc/argument`

The Combined Framework

In a recent work, Cabrio and Villata [67] propose to combine natural language techniques and Dung-like abstract argumentation to generate the arguments from natural language text and to evaluate this set of arguments to know which are the accepted ones, with the goal of supporting the participants in natural language debates (i.e. Debatepedia[22]). In particular, they adopt the TE approach, and in their experiments, they represent the TE relation extracted from natural language texts as a *support* relation in bipolar argumentation. In this section, we start from their observations, and we apply the combined framework proposed in [67] to this new scenario.

Let us consider the argument in Example 39 from the Wikipedia article "United States", and its revised versions in the last four years[23]:

Example 39.
In 2012: The land area of the contiguous United States is 2,959,064 square miles (7,663,941 km2).
In 2011: The land area of the contiguous United States is approximately 1,800 million acres (7,300,000 km2).
In 2010: The land area of the contiguous United States is approximately 1.9 billion acres (770 million hectares).
In 2009: The total land area of the contiguous United States is approximately 1.9 billion acres.

Several revisions have been carried out by different users during this four-year period, both to correct factual data concerning the U.S. surface, or to better specify them (e.g. providing the same value using alternative metric units). Following [67], we propose to take advantage of NLP techniques to automatically detect the relations among the revised versions of the same argument, to verify if the revisions done on the argument by a certain user at a certain point in time support the original argument (i.e. the user has rephrased the sentence to allow an easier comprehension of it, or has added more details), or attack it (i.e. the user has corrected some data, has deleted some details present in the previous version or has changed the semantics of the sentence providing a different viewpoint on the same content). Given the high similarities among the entailment and contradiction notions in TE and the support and attack relation in argumentation theory, we cast the described problem as a TE problem, where the T-H pair is a pair of revised arguments in two successive Wikipedia versions. We consider paraphrases as bidirectional entailment, and therefore to be annotated as a positive TE pair (i.e. support). Moreover, since the label *no entailment* includes both contradictions and pairs containing incomplete informational overlap (i.e. H is more informative than T), we consider both cases as *attacks*, since we want community managers to check the reliability of the corrected or deleted information. To build the T-H pairs required by the TE framework, for each argument we set the revised sentence as T and the original sentence as H, following the chronological sequence, since we want to verify if the more recent version entails or not the previous one, as shown in Example 40.

Example 40 (Continued)**.**
pair id=70.1 entailment=NO
T (Wiki12): The land area of the contiguous United States is 2,959,064 square miles (7,663,941 km2).
H (Wiki11): The land area of the contiguous United States is approximately 1,800 million acres (7,300,000 km2).

pair id=70.2 entailment=NO
T (Wiki11): The land area of the contiguous United States is approximately 1,800 million acres (7,300,000 km2).
H (Wiki10): The land area of the contiguous United States is approximately 1.9 billion acres (770 million hectares).

[22] `http://bit.ly/Dabatepedia`

[23] Since we are aware that Wikipedia versions are revised daily, we have picked our example from a random dump per year.

pair id=70.3 entailment=YES
T (Wiki10): The land area of the contiguous United States is approximately 1.9 billion acres (770 million hectares).
H (Wiki09): The total land area of the contiguous United States is approximately 1.9 billion acres.

On such pairs we apply a TE system, that automatically returns the set of arguments and the relations among them. The argumentation module starts from the couples of arguments provided by the TE module, and builds the complete argumentation framework involving such arguments. It is important to underline a main difference with respect to the approach of Cabrio and Villata [67]: here the argumentation frameworks resulting from the TE module represent a kind of *evolution* of the *same* argument during time in a specific Wikipedia article. From the argumentation point of view, we treat these arguments as separate instances of the same natural language argument giving them different names. Figure 5.7.a visualizes the argumentation framework of Example 40. This kind of representation of the natural language arguments and their evolution allows community managers to detect whether some arguments have been repeated in such a way that loops in the discussions can be avoided. The argumentation module, thus, is used here with a different aim from the previous approach [67]: it shows the *kind* of changes, i.e., positive and negative, that have been addressed on a particular argument, representing them using a graph-based structure.

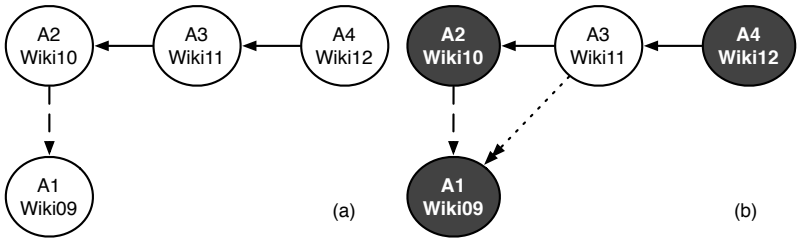


Figure 5.7: The bipolar argumentation framework resulting from Example 40.

The use of argumentation theory to discover the set of winning, i.e., acceptable, arguments in the framework could seem pointless, since we could assume that winning arguments are only those arguments appearing in the most recent version of the wiki page. However, this is not always the case. The introduction of the support relation in abstract argumentation theory [92] leads to the introduction of a number of *additional attacks* which are due to the presence of an attack and a support involving the same arguments. The additional attacks introduced in the literature are visualized in Figure 5.1, where dotted double arrows represent the additional attacks. For the formal properties of these attacks and a comparison among them, see Cayrol and Lagasquie-Schiex [92].

The introduction of additional attacks is a key feature of our argumentation module. It allows us to support community managers in detecting further possible attacks or supports among the arguments. In particular, given the arguments and their relations, the argumentation module builds the complete framework adding the additional attacks, and computes the extensions of the bipolar framework. An example of such kind of computation is shown in Figure 5.7.b where an additional attack is introduced. In this example, the set of accepted arguments would have been the same with or without the additional attack, but there are situations in which additional attacks make a difference. This means that the explicit attacks put forward by the users on a particular argument can then result in *implicit* additional attacks or supports to other arguments in the framework. Consider the arguments of Example 41. The resulting argumentation framework (see Figure 5.8) shows that argument $A1$ (*Wiki09*) is implicitly supported by argument $A4$ (*Wiki12*) since the attack of $A4$ (*Wiki12*) against $A3$ (*Wiki11*) leads to the introduction of an additional attack against $A2$

(*Wiki10*). The presence of this additional attack reinstates argument $A1$ (*Wiki09*) previously attacked by $A2$ (*Wiki10*). The two accepted arguments at the end are $\{A1, A4\}$.

Example 41.
pair id=7.1 entailment=NO
T (Wiki12): In December 2007, the United States entered its longest post-World War II recession, prompting the Bush Administration to enact multiple economic programs intended to preserve the country's financial system.
H (Wiki11): In December 2007, the United States entered the longest post-World War II recession, which included a housing market correction, a subprime mortgage crisis, soaring oil prices, and a declining dollar value.

pair id=7.2 entailment=YES
T (Wiki11): In December 2007, the United States entered the longest post-World War II recession, which included a housing market correction, a subprime mortgage crisis, soaring oil prices, and a declining dollar value.
H (Wiki10): In December 2007, the United States entered its longest post-World War II recession.

pair id=7.3 entailment=NO
T (Wiki10): In December 2007, the United States entered its longest post-World War II recession.
H (Wiki09): In December 2007, the United States entered the second-longest post-World War II recession, and his administration took more direct control of the economy, enacting multiple economic stimulus packages.

Finally, in this section we further enhance the framework proposed in [67] with a semantic machine readable representation of the argumentative discussions. We do not introduce yet another argumentation vocabulary, but we reuse the SIOC Argumentation module [193], focused on the fine-grained representation of discussions and argumentations in online communities.[24] The SIOC Argumentation model is grounded on DILIGENT [91] and IBIS[25] models.

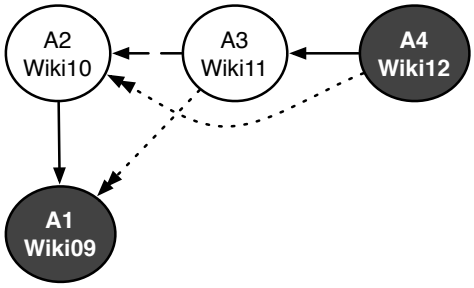


Figure 5.8: The bipolar argumentation framework resulting from Example 41.

We extend the SIOC Argumentation vocabulary with two new properties `sioc_arg:challengesArg` and `sioc_arg:supportsArg` whose range and domain are `sioc_arg:Argument`. These properties represent challenges and supports from arguments to arguments, as required in abstract argumentation theory.[26] This needs to be done since in SIOC Argumentation challenges and supports are addressed from arguments towards `sioc_arg:Statement` only. Figure 5.9.a shows a sample of the semantic representation of Example 31 and 32 where *contradiction* is represented through `sioc_arg:challengesArg`, and *entailment* is represented through `sioc_arg:supportsArg`.

[24] For an overview of the argumentation models in the Social Semantic Web, see [286].
[25] `http://purl.org/ibis`
[26] The extended vocabulary can be downloaded at `http://bit.ly/SIOC_Argumentation`

```
EXAMPLE OF CONTRADICTION
<http://example.org/jako/pair1t> rdf:type sioc_arg:Argument ;

     sioc:content "It was reported that Jackson had
            offered to buy the bones of Joseph Merrick
            (the elephant man) and although untrue,
            Jackson did not deny the story." ;

     sioc_arg:challengesArg <http://example.org/jako/pair1h> .

<http://example.org/jako/pair1h> rdf:type sioc_arg:Argument ;

        sioc:content "Later it was reported that Jackson
            bought the bones of The Elephant Man." .

EXAMPLE OF ENTAILMENT
<http://example.org/jako/pair2t> rdf:type sioc_arg:Argument ;

        sioc:content "Jackson had three sisters: Rebbie,
            La Toya, and Janet, and six brothers: Jackie,
            Tito, Jermaine, Marlon, Brandon (Marlon's twin
            brother, who died shortly after birth) and
            Randy." ;

       sioc_arg:supportsArg <http://example.org/jako/pair2h> .

<http://example.org/jako/pair2h> rdf:type sioc_arg:Argument ;

        sioc:content "Jackson's siblings are Rebbie, Jackie,
            Tito, Jermaine, La Toya, Marlon, Randy and
            Janet." .
```

```
PREFIX sioc_arg:<http://rdfs.org/sioc/argument#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX sioc:<http://rdfs.org/sioc/ns#>

SELECT ?a1 ?c1 WHERE {
   ?a1 a sioc_arg:Argument .
   ?a2 a sioc_arg:Argument .
   ?a1 sioc_arg:challengesArg ?a2 .
   ?a1 sioc:content ?c1 .
   ?a2 sioc:content ?c2
   FILTER regex(str(?c2),"crisis")
}
```

QUERY RESULT

```
T: "In December 2007, the United States entered its longest post–World
War II recession, prompting the Bush Administration to enact multiple
economic programs intended to preserve the country's financial system."
ATTACKS
H: "In December 2007, the United States entered the longest post–World
War II recession, which included a housing market correction, a subprime
mortgage crisis, soaring oil prices, and a declining dollar value."

T: "Bush entered office with the Dow Jones Industrial Average at 10,587,
and the average peaked in October 2007 at over 14,000."
ATTACKS
H: "The Dow Jones Industrial Average peaked in October 2007 at about
14,000, 30 percent above its level in January 2001, before the subsequent
economic crisis wiped out all the gains and more."
```

Figure 5.9: (a) Sample of the discussions in RDF, (b) Example of SPARQL query.

The semantic version of the argumentative discussions can further be used by community managers to detect insightful meta-information about the discussions themselves. For instance, given the RDF data set being stored in a datastore with SPARQL endpoint, the community manager can raise a query like the one in Figure 5.9.b. This query retrieves all those arguments which attack another argument having in the content the word "crisis". This simple example shows how the semantic annotation of argumentative discussions may be useful to discover in an automatic way those information which are difficult to be highlighted by a human user.

Experimental Setting

As a case study to experiment our framework we select the Wikipedia revision history. We first describe the creation of the data set, then we discuss the TE system we used, and we report on obtained results.

Data Set. We create a data set to evaluate the use of TE to generate the arguments following the methodology detailed in [66]. We start from two dumps of the English Wikipedia (*Wiki 09* dated 6.03.2009, and *Wiki 10* dated 12.03.2010), and we focus on the five most revised pages[27] at that time (i.e. George W. Bush, United States, Michael Jackson, Britney Spears, and World War II). We then follow their yearly evolution up to now, considering how they have been revised in the next Wikipedia versions (*Wiki 11* dated 9.07.2011, and *Wiki 12* dated 6.12.2012).

After extracting plain text from the above mentioned pages, for both *Wiki 09* and *Wiki 10* each document has been sentence-splitted, and the sentences of the two versions have been automatically aligned to create pairs. Then, to measure the similarity between the sentences in each pair, following [66] we adopted the *Position Independent Word Error Rate (PER)*, i.e. a metric based on the calculation of the number of words

[27] http://bit.ly/WikipediaMostRevisedPages

which differ between a pair of sentences. For our task we extracted only pairs composed by sentences where major editing was carried out (*0.2 < PER < 0.6*), but still describe the same event.[28] For each pair of extracted sentences, we create the TE pairs setting the revised sentence (from *Wiki 10*) as T and the original sentence (from *Wiki 09*) as H. Starting from such pairs composed by the same revised argument, we checked in the more recent Wikipedia versions (i.e. *Wiki 11* and *Wiki 12*) if such arguments have been further modified. If that was the case, we created another T-H pair based on the same assumptions as before, i.e. setting the revised sentence as the T and the older sentence as the H (see Example 40). Such pairs have then been annotated with respect to the TE relation (i.e. *YES/NO entailment*), following the criteria defined and applied by the organizers of the Recognizing Textual Entailment Challenges (RTE)[29] for the two-way judgment task.

As a result of the first step (i.e. extraction of the revised arguments in *Wiki 09* and *Wiki 10*) we collected 280 T-H pairs, while after applying the procedure on the same arguments in *Wiki 11* and *Wiki 12* the total number of collected pairs is 452. To carry out our experiments, we randomly divided such pairs into training set (114 entailment, 114 no entailment pairs), and test set (101 entailment, 123 no entailment pairs). The pairs collected for the test set are provided in their unlabeled form as input to the TE system. To correctly train the TE system we balanced the data set with respect to the percentage of yes/no judgments. In Wikipedia, the actual distribution of attacks and supports among revisions of the same sentence is slightly unbalanced since generally users edit a sentence to add different information or correct it, with respect to a simple reformulation.[30]

To assess the validity of the annotation task and the reliability of the obtained data set, the same annotation task has been independently carried out also by a second annotator, so as to compute inter-annotator agreement. It has been calculated on a sample of 140 argument pairs (randomly extracted).

The inter-annotator agreement results in $\kappa = 0.82$. As a rule of thumb, this is a satisfactory agreement, therefore we consider these annotated data sets as the *goldstandard*[31], i.e. the reference data set to which the performances of our combined system are compared. As introduced before, the goldstandard pairs have then been further translated into RDF using SIOC Argumentation.[32]

TE System. To detect which kind of relation underlies each couple of arguments, we use the EDITS system (Edit Distance Textual Entailment Suite) version 3.0, an open-source software package for RTE[33] [189]. EDITS implements a distance-based framework which assumes that the probability of an entailment relation between a given T-H pair is inversely proportional to the distance between T and H (i.e. the higher the distance, the lower is the probability of entailment).[34] Within this framework the system implements different approaches to distance computation, i.e. both edit distance and similarity algorithms. Each algorithm returns a normalized distance score (a number between 0 and 1). At a training stage, distance scores calculated over annotated T-H pairs are used to estimate a threshold that best separates positive from negative examples, that is then used at a test stage to assign a judgment and a confidence score to each test pair.

[28] A different extraction methodology has been proposed in [330].

[29] http://www.nist.gov/tac/2010/RTE/

[30] As introduced before, we set a threshold in our extraction procedure to filter out all the minor revisions, concerning typos or grammatical mistakes corrections.

[31] The dataset is available at http://www-sop.inria.fr/NoDE/

[32] The obtained data set is downloadable at http://www-sop.inria.fr/NoDE/

[33] http://edits.fbk.eu/

[34] In previous RTE challenges, EDITS always ranked among the 5 best participating systems out of an average of 25 systems, and is one of the two RTE systems available as open source http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool

Table 5.6: Systems performances on Wikipedia data set

		Train			Test		
EDITS configurations	*rel*	*Precision*	*Recall*	*Accuracy*	*Precision*	*Recall*	*Accuracy*
WordOverlap	*yes*	0.83	0.82	**0.83**	0.83	0.82	**0.78**
	no	0.76	0.73		0.79	0.82	
CosineSimilarity	*yes*	0.58	0.89	0.63	0.52	0.87	0.58
	no	0.77	0.37		0.76	0.34	

Evaluation. To evaluate our framework, we carry out a two-step evaluation: first, we assess the performances of EDITS to correctly assign the *entailment* and the *no entailment* relations to the pairs of arguments on the Wikipedia data set. Then, we evaluate how much such performances impact on the application of the argumentation theory module, i.e. how much a wrong assignment of a relation to a pair of arguments is propagated in the argumentation framework. For the first evaluation, we run EDITS on the Wikipedia training set to learn the model, and we test it on the test set. In the configurations of EDITS we experimented, the distance entailment engine applies *cosine similarity* and *word overlap* as the core distance algorithms. In both cases, distance is calculated on lemmas, and a stopword list is defined to have no distance value between stopwords. Obtained results are reported in Table 5.6. Due to the specificity of our data set (i.e. it is composed by revisions of arguments), *word overlap* algorithm outperforms *cosine similarity* since there is high similarity between revised and original arguments (in most of the positive examples the two sentences are very close, or there is an almost perfect inclusion of H in T). For the same reason, obtained results are higher than in [67], and than the results obtained on average in RTE challenges. For these runs, we use the system off-the-shelf, applying its basic configuration. As future work, we plan to fully exploit EDITS features, integrating background and linguistic knowledge in the form of entailment rules, and to calculate the distance between T and H based on their syntactic structure.

As a second step in our evaluation phase, we consider the impact of EDITS performances (obtained using word overlap, since it provided the best results) on the acceptability of the arguments, i.e. how much a wrong assignment of a relation to a pair of arguments affects the acceptability of the arguments in the argumentation framework. We use admissibility-based semantics [132] to identify the accepted arguments both on the correct argumentation frameworks of each Wikipedia revised argument (where entailment/contradiction relations are correctly assigned, i.e. the goldstandard), and on the frameworks generated assigning the relations resulted from the TE system judgments. The precision of the combined approach we propose in the identification of the accepted arguments is on average 0.90 (i.e. arguments accepted by the combined system and by the goldstandard w.r.t. a certain Wikipedia revised argument), and the recall is 0.92 (i.e. arguments accepted in the goldstandard and retrieved as accepted by the combined system). The F-measure (i.e. the harmonic mean of precision and recall) is 0.91, meaning that the TE system mistakes in relation assignment propagate in the argumentation framework, but results are still satisfying and foster further research in this direction. For this feasibility study, we use four Wikipedia versions, so the resulting AFs are generally composed by four couples of arguments connected by attacks or supports. Reduced AFs are produced when a certain argument is not revised in every Wikipedia version we considered, or when an argument is deleted in more recent versions. Using more revised versions will allow us to generate even more complex argumentation graphs.

In this section, we presented a framework to support community managers in managing argumentative discussions on wiki-like platforms. In particular, our approach proposes to automatically detect the natural language arguments and the relations among them, i.e., support or challenges, and then to organize the detected arguments in bipolar argumentation frameworks. This kind of representation helps community managers to understand the overall structure of the discussions and which are the winning arguments. Moreover, the generated data set is translated in RDF using an extension of the SIOC Argumentation vocabulary such that the discussions can be queried using SPARQL in order to discover further insightful information. The experimental evaluation shows that in 85% of the cases, the proposed approach correctly detects the accepted arguments. SIOC[35] allows to connect the arguments to the users who propose them. This is important in online communities because it allows to evaluate the arguments depending on the expertise of their sources. In this section, we do not represent users neither in the argumentation frameworks nor in the RDF representation of the discussions, and this is left as future work. Moreover, we plan to move from the crisp evaluation of the arguments' acceptability towards a more flexible evaluation where the expertise of the users proposing the arguments plays a role. As future work on the NLP side, we consider experimenting a TE system carrying out a three-way judgment task (i.e. *entailment*, *contradiction* and *unknown*), to allow for a finer-grained classification of non entailment pairs (i.e. to separate when T contradicts H, from when H is more informative than T).

5.5 Argument Mining on Social Media

Argumentation has come to be increasingly central as a main study within Artificial Intelligence, due to its ability to conjugate representational needs with user-related cognitive models and computational models for automated reasoning. An important source of data for many of the disciplines interested in such studies is the Web, and social media in particular. Newspapers, microblogs, online debate platforms and social networks provide an heterogeneous flow of information where natural language arguments can be identified and analyzed. The availability of such data, together with the advances in Natural Language Processing and Machine Learning, supported the rise of a new research area called *argument mining*, whose main goal is the automated extraction of natural language arguments and their relations from generic textual corpora, with the final purpose of providing machine-processable data for computational models of argument.

Despite the increasing amount of argument mining approaches [206], none of them has tackled the challenge of extracting arguments and their relations on social media like Twitter or Facebook. Such a kind of natural language arguments raises further issues in addition to the standard problems faced by argument mining approaches typically dealing with newspapers, novels or legal texts: messages from Twitter are squeezed, noisy and often unstructured. More specifically, the following issues have to be considered: *i)* the 140-characters limit forces users to express their ideas very succinctly; *ii)* the quality of the language in Twitter is deteriorated, including a lot of variants in spelling, mistakes and abbreviations, and *iii)* Twitter's API filters tweets on hashtags but cannot retrieve all the replies to these tweets if they do not contain the same hashtags.

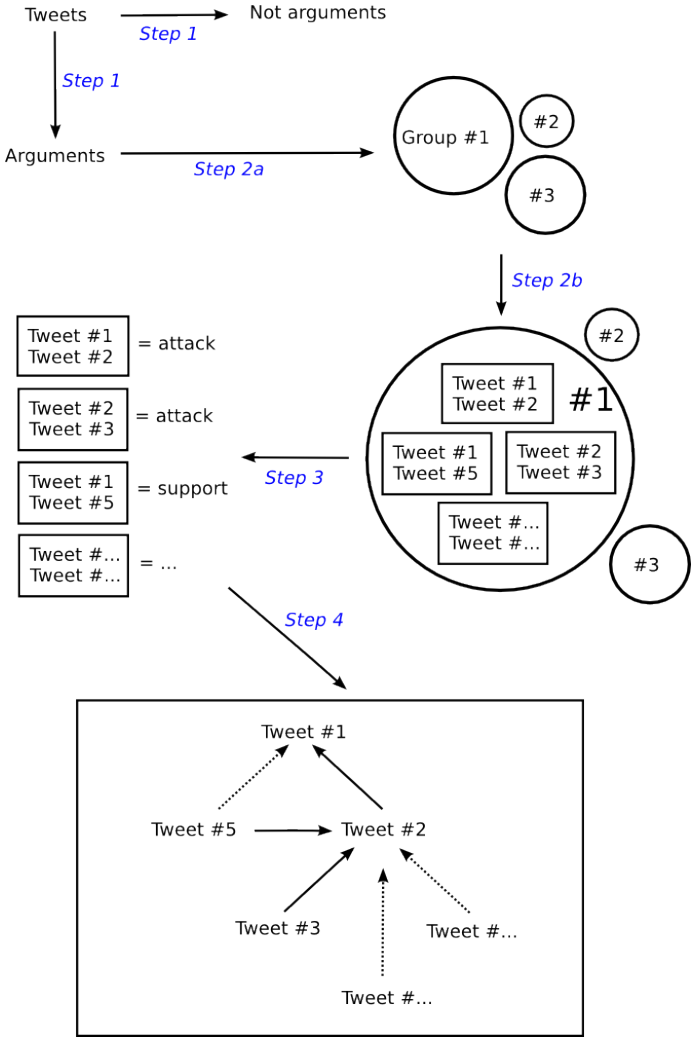
In this section, we provide a preliminary answer to the following research question: *how to extract the arguments and predict the relations among them on Twitter data?* and we highlight the open challenges still to be addressed. We consider both the two main stages in the typical argument mining pipeline, from the unstructured natural language documents towards structured data: we first detect arguments within the natural language texts from Twitter, the retrieved arguments will thus represent the nodes in the final argument

[35] `http://sioc-project.org`

graph returned by the system, and second, we predict what are the relations, i.e., *attack* or *support*, holding between the arguments identified in the first stage.

The main advantage of our approach is that it provides a whole argument mining pipeline to analyze flows of tweets, allowing for the application of reasoning techniques over the output structured data, like the identification of the set of widely accepted arguments or trends analysis. However, being it an ongoing work, we highlight in this section both positive and negative results in applying argument mining on Twitter data, analyzing solutions and potential alternatives to be explored.

Figure 5.10: Pipeline architecture



Argument Mining on Twitter

The argument mining pipeline we propose, visualized in Figure 5.10, is composed of four main steps, that consist in: *i)* separating tweet-arguments from non-argument tweets; *ii)* grouping tweet-arguments

discussing about the same issue, and create pairs of arguments; *iii)* predicting the relations of *attack* and *support* among the tweets in the pairs; and *iv)* building argumentation graphs.

First of all, we need to clarify what we mean by *argument* in this section: an argument gives a reason to support a claim that is questionable, or open to doubt. In the computational models of argument field, an argument is made of three components: the *premises* representing the reason, a *conclusion* which is the supported claim, and a *relation* showing how the premises lead to this conclusion. Facing the issue of dealing with Twitter data, i.e., dealing with textual arguments of length inferior or equal to 140 characters, we (almost) never find such a kind of complete structure of the arguments. We have thus labeled as arguments all those text snippets providing a portion of a standard argument structure, e.g., opinions under the form of claim, data like in the Toulmin model [302], or persuasive conclusions. Future work includes the "composition" of such elements to build a single well-structured argument. Second, it is worth noticing that the support and the attack relations are not symmetric: we considered the temporal dimension to decide the direction of these relations, i.e., a tweet that is proposed at time $t+1$ attacks (resp. supports) a tweet which has been provided at time t. In the following, each step of the pipeline is described in detail, together with the experimented approach, and the obtained results of this ongoing work.

Dataset. Up to our knowledge, DART [61] is the only existing dataset of arguments and their relations on Twitter, therefore it has been chosen to test our pipeline. It is composed of:

(a) 4000 tweets annotated as argument/not argument: 1000 tweets for each of the following 4 topics: the letter to Iran written by 47 senators on 10/03/2015; the referendum in Greece for or against Greece leaving European Union on 10/07/2015; the release of Apple Watch on 10/03/2015; the airing of episode 4 (season 5) of the serie Game of Thrones on 4/05/2015. A tweet is annotated as argument if it contains an opinion or factual information, or if it is a claim expressed as question (rhetorical questions, attempts to persuade, containing sarcasms/irony). The argument annotation task is carried out on a single tweet and not on subparts of it.

A text containing an opinion is considered as an argument. For example, in the following tweet the opinion of the author is clearly expressed in the second sentence (i.e., *I won't be running out to get one*):

> *RT @mariofraioli: What will #AppleWatch mean for runners? I can't speak for everyone, but I won't be running out to get one. Will you? http://t.co/xBpj0HWK*

We consider as arguments also claims expressed as questions (either rhetorical questions, attempts to persuade, containing sarcasm or irony), as in the following example:

> *RT @GrnEyedMandy: What next Republicans? You going to send North Korea a love letter too? #47Traitors*

or:

> *Perhaps Apple can start an organ harvesting program. Because I only need one kidney, right? #iPadPro #AppleTV #AppleWatch*

Tweets containing factual information are annotated as arguments, given that they can be considered as premises or conclusions. For example:

RT @HeathWallace: You can already buy a fake #AppleWatch in China http://t.co/WpHEDqYuUC via @cnnnews @mr_gadget http://t.co/WhcMKuM

Defining the amount of world knowledge needed to determine whether a text is a fact or an opinion when it contains unknown acronyms and abbreviations can be pretty tough. Consider the following tweet:

RT @SaysSheToday: The Dixie Chicks were attacked just for using 1A right to say they were ashamed of GWB. They didn't commit treason like the #47Senators

where the mentioned entities *The Dixie Chicks*, *GWB*, and *1A right* are strictly linked to the US politics, and hardly interpreted by people out of the US politics matters. In this case, annotators are asked to suppose that the mentioned entities exist, and focus on the phrasing of the tweets.

However, if tweets contain pronouns only (preventing the understanding of the text), we consider such tweets as not "self-contained" , and thus non arguments. It can be the case of replies, as in the following example, in which the pronoun *he* is not referenced anywhere in the tweet.

@FakeGhostPirate @GameOfThrones He is the one true King after all ;)

For tweets containing an advertisement to push into visiting a web page, if an opinion or factual information is also present, then the tweet is considered as an argument, otherwise it is not. Consider the following example:

RT @NewAppleDevice: Apple's smartwatch can be a games platform and here's why http://t.co/uIMGDyw08I

It contains factual information that can be understood even without visiting the link. On the contrary, the following tweet is not an argument, given that it does not convey an independent message while excluding the link:

For all #business students discussing #AppleWatch this morning. Give it a test drive thanks to @UsVsTh3m: http://t.co/x2bGc9j1Gl.

(b) 2181 tweet-arguments on the Apple Watch release classified in 7 categories (i.e. *features (F)*, *price (P)*, *look (L)*, *buying announcement (B)*, *advertisement (A)*, *forecast on the product success (S)*, *news (N)*, *others (O)*) (see Table 5.9). Moreover, the tweets contained in the category *features* have been grouped in the following more fine-grained categories: *health*, *innovation*, *battery*.

(c) 1891 pairs of tweet-arguments of the categories: *price*, *health*, *look*, *predictions* annotated with the following relations: *support* (446), *attack* (122), *unknown* (1323). After a first annotation round to test the guidelines provided in [70], we realized that a few additional instructions should be added with the aim to consider the specificity of the Twitter scenario. The instructions we introduced are as follows:

If both Tweet-A and Tweet-B in a pair are factual tweets, and they are related to the same issue, the pair must be annotated as *support*, as in:

Tweet-A: *.@AirStripmHealth + #AppleWatch provides HIPPA compliant capabilities for physicians, mothers, babies, and more #AppleEvent*

Tweet-B: *accessible heart rate monitors and opinions on that #iWatch #apple #accessibility #ios* `https://t.co/ySYM8dk0Pf` *via @audioBoom*

If both Tweet-A and Tweet-B in a pair are opinion tweets, and they are related to the same issue, the pair must be annotated as *support*, as in:

Tweet-A: *Think of how much other stuff you can buy with the money you spend on an #AppleWatch*

Tweet-B: *#AppleWatch Tempting, but not convinced. #appletv Yes. #iPhone6sPlus No plan to upgrade #iPadPro little high price, wait & watch*

If Tweet-B is a factual tweet, and Tweet-A is an opinion on the same issue, the pair must be annotated as *support*, as in:

Tweet-A: *Wow. Your vitals on your iwatch. That's bonkers. #AppleEvent*

Tweet-B: *accessible heart rate monitors and opinions on that #iWatch #apple #accessibility #ios* `https://t.co/ySYM8dk0Pf` *via @audioBoom*

If Tweet-A is a factual tweet, and Tweet-B expresses someone's wishes to buy the product or an opinion about it, the pair must be annotated as *unknown*, as in:

Tweet-A: *Mom can listen to baby's heart rate with #AppleWatch #airstrip*

Tweet-B:*Wow!!! Look at what the #Ap, pleWatch can do for #doctors that's amazing! Seeing their vitals? I just got chills! In a good way #AppleEvent*

Concerning the annotation of the arguments/non arguments, in the reconciliation phase among the three students annotators, the label that was annotated by at least 2 annotators out of 3 was chosen (majority voting mechanism). If all the annotators disagree or if more than one annotator labels the tweet as unknown, then such tweet is discarded. The inter-annotator agreement has been calculated between the expert annotators and the reconciled student annotations on 250 tweets of the first batch, resulting in $\alpha_{47traitors} = 0.81$ (Krippendorff's α handles missing values, the label "unknown" in our case). Concerning the pair annotation with the support/attack/unknown relations, the inter-annotator agreement has been calculated on 99 pairs (33 pairs randomly extracted from each of the three first topics), resulting in Krippendorff $\alpha = 0.67$.

Step 1: Argument identification.

The first task in our pipeline is the binary classification of tweets as argument/non argument. To train a generic, domain-independent argument detector, we separate the training, validation and test data according to the topics of dataset (a) to avoid overfitting. We train and validate on the first three topics, and we test on the Apple Watch dataset (Table 5.7 provides some statistics on the data).We ignore tweets classified as unknown. We use 3-fold cross-validation (we alternately train the model on the tweets of the first two topics and leave the third topic out as a validation set) with randomized hyperparameter search [40].[36] Because

[36]A randomized hyperparameter search samples parameter settings a fixed number of times and has been found to be more effective in high-dimensional spaces than exhaustive search.

Dataset	# tweet-arg.	not-arg.	unknown	total
Training set	2079	829	92	3000
Test set	623	352	25	1000

Table 5.7: Statistics of dataset (a)

Approach	Average F1
baseline	0.64
baseline + tokens	0.66
baseline + tokens + bigrams tokens	0.67

Table 5.8: Validation of the model and feature use

the classes are unbalanced and the balance is not necessarily the same across all datasets, the training phase weights the errors inversely proportional to class frequencies.

As baseline, we use raw character counts as features (causing smileys, capital letters, punctuation marks to influence the model). Then, tweets have been tokenized with Twokenize[37] and annotated with their PoS applying Stanford POS tagger. POS tags are then used as features, as well as bigrams of tags. As a baseline model, we train a logistic regression model[38] on these features only.

We also augment features with normalized tokens and bigrams of tokens, and this effectively improves over the baseline (see Table 5.8). The best model (Logistic regression, L2-penalized with $\lambda = 100$) is obtained by using all the features and re-training on the 3 folds. It yields an F1-score of 0.78 over the test set, that can be considered as satisfactory. The difference between the average F1-score over the validation set (see Table 5.8) and the F1 over the test set is due to the addition of the tweets of the validation set (around 1000 additional tweets) for training the final model.

Step 2: Pairs creation.

Once we are able to identify tweet-argument, we create pairs of them to predict the relations among them. Given a stream of tweets, it would be impossible to apply a naive approach comparing all the pairs of tweets, since this would lead to the creation of numerous unrelated pairs.

To deal with this issue, we firstly tested the solution of clustering the tweets into *sub-topics*, and then create pairs from these sub-topics. The major problem that we faced is the difficulty of automatically finding meaningful sub-topics. We tested both Latent Dirichlet Allocation[39] [50] and more powerful models such as Correlated Topic Models[40] [49], but the interpretability of the clusters did not improve [101].

[37] `http://www.cs.cmu.edu/~ark/TweetNLP/`

[38] Like all regression analyses, the logistic regression is a predictive analysis. It is used to describe data and to measure the relationship between one dependent variable and one or more independent variables by estimating probabilities using a logistic function, i.e., the cumulative logistic distribution.

[39] Latent Dirichlet allocation is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

[40] Correlated Topic Models use a more flexible distribution for the topic proportions that allows for covariance structure among the components. This gives a more realistic model of latent topic structure where the presence of one latent topic may be correlated with the presence of another.

	O	A	B	F	L	N	P	S
#	720	175	370	619	205	65	189	112

Table 5.9: Statistics on dataset (b), # tweets

	F	L	P	S
average F1-score (train set)	0.36	0.57	0.60	0.15
F1-score (test set)	0.56	0.58	0.60	0.00

Table 5.10: Classification results (step 2)

Instead, since we have classified goldstandard data for Apple Watch (dataset (b), see Table 5.9), we decided to focus on this topic only, and turn the clustering problem into a classification problem. Another possibility would have been to tune the hyperparameters before applying the clustering algorithms to retrieve the annotated categories, but given the small size of the goldstandard, we could not explore that direction further.

In particular, we focus on categories F (features), L (look), P (price) and S (predictions about the success of the product) because they contain the most interesting tweets. We use the same features and same hyper-parameters selection scheme as in step 1. The training set contains 2031 tweets, and the test set contains 150 tweets. The 3 folds are randomly created across all the training set, and we take the average of all the macro F1-scores on all the folds to select the best model. We use regularized logistic regression and the results obtained by the best model (L1-penalized with $\lambda = 100$) are reported in Table 5.10 for each category, averaged over all the folds. As can be observed, some categories are harder to predict than others, but the performance on the easy classes (F, L, P here) are quite satisfactory. A paraphrase detection tool could be added at this step to deduplicate similar tweets and give more weights to the arguments that are often used in subsequent steps.

Step 3: Relation detection.

Given the pairs of tweet-arguments returned by step 2, the next step consists in predicting the relation holding between the tweets in a pair. Dataset (c) contains $\sim$600 tweets each for *look*, *price* and *health* categories of the Apple Watch: we put pairs concerning the product price in the test set, whereas all the other tweets are in the training set. An additional validation set contains 100 tweets on the user predictions on the product success.

Given the closeness of the task with textual entailment [70], we decide to explore first a prediction of the support and attack relations using the Excitement Open Platform (EOP)[41] for recognizing textual entailment. The intuition is to consider the support relation as an entailment, and the attack relation as a contradiction, following the approach in Cabrio and Villata [71].

In addition, following the same guidelines proposed by [70], pairs are also annotated according to the Recognizing Textual Entailment (RTE) framework, i.e., pairs linked by a support relation as *entailment/non-entailment*, and pairs linked by an attack relation as *contradiction/non-contradiction*.

[41] http://hltfbk.github.io/Excitement-Open-Platform/

Model	EOP (MaxEnt)	Neural model
F1-score Support	0.17	0.20
F1-score Attack	0.0	0.16

Table 5.11: Comparing the two models

However, given the specificity of Twitter data and the fact that predicting support and attack relations is not the same as recognizing entailment, results were far from being satisfying (see Table 5.11), also due to the huge number of unrelated pairs (tagged as unknown in Dataset (c)). Then we decided to implement a neural sequence classifier inspired by [273]. We encode the tokens as precomputed GloVe embeddings[42] [257] of size 200. When a token does not have an embedding, we generate a random embedding according to a multivariate normal distribution with empirical mean and variance of existing embeddings.

Such a neural classifier is an encoder-decoder architecture with two distinct Long Short-Term Memory networks[43] (LSTM) [170], where we pass the last hidden-state of the first LSTM to initialize the second. The probabilities over the 3 categories are given by a softmax function, i.e., a function which takes as input a C-dimensional vector z and outputs a C-dimensional vector y of real values between 0 and 1, at the output layer of the second LSTM at the last pass. Our objective is cross entropy, and we oversample the attack and support categories so that the probability of drawing a tweet from a category is uniform on the three categories. We use Stochastic Gradient Descent with Adam[44] [188] to optimize. We periodically test our model against the validation set, and stop the training when the validation error stops improving. We select the best performing model on the validation set. However, also in this case, results are not satisfying (see Table 5.11).

We realize that such classification step on Twitter is pretty hard, even for human. As an example, consider the following pair:

T1: *Can't believe the designers of #AppleWatch didn't present a better shaped watch. It's still too clunky looking & could've been more sleek.*
T2: *@APPLEOFFICIAL amazing product updates. Apple TV looks great. BUT! Please make a bigger iWatch! Not buying it until it's way bigger.*

On the one hand, the tweets agree in that the watch is not properly sized. On the other hand, they disagree since one user finds it too big and the other one too small, which are opposite viewpoints.

The neural model is more promising because it can be easily used in a semi-supervised settings, but the lack of a large-sized corpus is a huge hurdle for training such a model (however, there is a huge amount of data in the DART dataset that has not been labeled yet, for which an annotation effort should be considered).

[42] GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

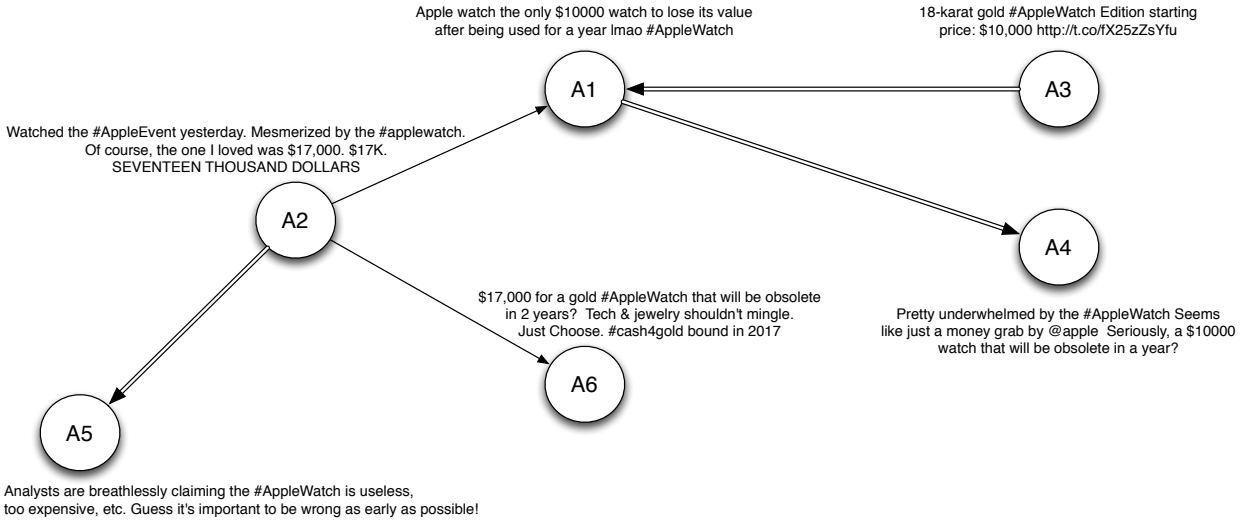
[43] Long Short-Term Memory networks are a special kind of Recurrent Neural Networks, capable of learning long-term dependencies.

[44] Adam is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments.

Step 4: Graph building

We can now build an *argument graph* whose nodes are the arguments and whose edges are the predicted relations (supports/ attacks). An example of such a graph is visualized in Figure 5.11, where an extract of the tweets for the iWatch topic is presented. It is easy to note that such a kind of visualization allows for a deeper understanding of the ongoing Twitter discussion, and would provide a valuable support for social media content analysis.

Figure 5.11: Example of argumentation graph (where single edges represent attack and double ones represent support) resulting from the identified arguments and predicted relations for the iWatch topic.



The last step of the pipeline consists in applying argumentation semantics to identify the set(s) of accepted arguments. Several systems can be adopted to perform such a computation in a scalable way, as those participating to the ICCMA challenge [300]. In our framework, we used the ASPARTIX-D system[45], after the flattening of the bipolar argumentation framework to an abstract Dung-like argumentation framework, as done in [70]. This step returns the set of acceptable arguments such that the different (coherent) viewpoints expressed through the tweets are highlighted, as well as the identifiable attack points in the stream.

Some considerations can be drawn about the resulting graphs. First of all, graphs are, differently from [73] for instance, rather sparse, meaning that they do not present a star structure. They are more like a set of subgraphs connected with each other, where each subgraph concerns a different sub-issue of the general topic, i.e., the price of the Hermes iWatch band inside the *Price* issue of the iWatch topic. This is a specificity of Twitter discussions being them a continuous stream of messages. Second, as for the case of the debates extracted in [73], no cycle is present.

Given the unsatisfactory results obtained for the relation prediction task given the huge difference in the selected topics of the DART dataset, we decided to go a step further and address the following sub-questions, starting from the argument detection component, that arise in the context of social media: *i)* how to distinguish factual arguments from opinions? *ii)* how to automatically detect the source of factual arguments? To answer these questions, we extend and annotate a dataset of tweets extracted from the streams about the Grexit and the Brexit news. To address the first task of argument detection, we apply supervised classi-

[45]https://ddll.inf.tu-dresden.de/web/Sarah_Alice_Gaggl/ASPARTIX-D

fication to separate *argument-tweets* from non-argumentative ones. By considering only argument-tweets, in the second step we apply again a supervised classifier to recognize tweets reporting factual information from those containing opinions only. Finally, we detect, for all those arguments recognized as factual in the previous step, what is the source of such information (e.g., the CNN), relying on the type of the Named Entities recognized in the tweets. The last two steps represent new tasks in the argument mining research field, of particular importance in social media applications.

From arguments to facts and sources

Dataset. The only available resource of annotated tweets for argument mining is DART [60].[46] From the highly heterogeneous topics contained in such resource (i.e. the letter to Iran written by 47 U.S. senators; the referendum for or against Greece leaving the EU; the release of Apple iWatch; the airing of the 4th episode of the 5th season of the TV series Game of Thrones), and considering the fact that tweets discussing a political topic generally have a more developed argumentative structure than tweets commenting on a product release, we decided to select for our experiments the subset of the DART dataset on the thread *#Grexit* (987 tweets). Then, following the same methodology described in [60], we have extended such dataset collecting 900 tweets from the thread on *#Brexit*. From the original thread, we filtered away retweets, accounts with a bot probability $>$0.5 [120], and almost identical tweets (Jaccard distance, empirically evaluated threshold). Given that tweets in DART are already annotated for task 1 (argument/non-argument, see Section 5.5), two annotators carried out the same task on the newly extracted data. Moreover, the same annotators annotated both datasets (Grexit/Brexit) for the other two tasks of our experiments, i.e. *i)* given the argument tweets, annotation of tweets as either containing factual information or opinions (see Section 5.5), and *ii)* given factual argument tweets, annotate their source when explicitly cited (see Section 5.5). Tables 5.12, 5.13 and 5.14 contain statistical information on the datasets.

Inter annotator agreement (IAA) [86] between the two annotators has been calculated for the three annotation tasks, resulting in κ=0.767 on the first task (calculated on 100 tweets), κ=0.727 on the second task (on 80 tweets), and Dice=0.84 [125][47] on the third task (on the whole dataset). More specifically, to compute IAA, we sampled the data applying the same strategy: for the first task, we randomly selected 10% of the tweets of the Grexit dataset (our training set); for task 2, again we randomly selected 10% of the tweets annotated as argument in the previous annotation step; for task 3, given the small size of the dataset, both annotators annotated the whole corpus.

dataset	# argument	# non-arg	total
Brexit	713	187	900
Grexit	746	241	987
total	1459	428	1887

Table 5.12: Dataset for task 1: argument detection

[46] Annotated data are available upon request to the authors.

[47] Dice is used instead of κ to account for partial agreement on the set of sources detected in the tweets.

dataset	# factual arg.	# opinion	total
Brexit	138	575	713
Grexit	230	516	746
total	368	1091	1459

Table 5.13: Dataset for task 2: factual arguments vs opinions classification

dataset	# arg. with source cit.	# arg. without source cit.	total
Brexit	40	98	138
Grexit	79	151	230
total	119	249	368

Table 5.14: Dataset for task 3: source identification

Classification algorithms. We tested Logistic Regression (LR) and Random Forest (RF) classification algorithms, relying on the *scikit-learn* tool suite[48]. For the learning methods, we have used a Grid Search (exhaustive) through a set of predefined hyper-parameters to find the best performing ones (the goal of our work is not to optimize the classification performance but to provide a preliminary investigation on new tasks in argument mining over Twitter data). We extract argument-level features from the dataset of tweets (following [319]), that we group into the following categories:

- *Lexical (L):* unigram, bigram, WordNet verb synsets;
- *Twitter-specific (T):* punctuation, emoticons;
- *Syntactic/Semantic (S):* we have two versions of dependency relations as features, one being the original form, the other generalizing a word to its POS tag in turn. We also use the syntactic tree of the tweets as feature. We apply the Stanford parser [214] to obtain parse trees and dependency relations;
- *Sentiment (SE):* we extract the sentiment from the tweets with the Alchemy API[49], the sentiment analysis feature of IBM's Semantic Text Analysis API. It returns a polarity label (positive, negative or neutral) and a polarity score between -1 (totally negative) and 1 (totally positive).

As baselines we consider both LR and RF algorithms with a set of basic features (i.e., lexical).

Task 1: Argument detection

The task consists in classifying a tweet as being an argument or not. We consider as arguments all those text snippets providing a portion of a standard argument structure, i.e., opinions under the form of claims, facts mirroring the data in the Toulmin model of argument [302], or persuasive claims, following the definition

[48] http://scikit-learn.org/

[49] https://www.ibm.com/watson/alchemy-api.html

of argument tweet provided in [60, 59]. Our dataset contains 746 argument tweets and 241 non-argument tweets for Grexit (that we use as training set), and 713 argument tweets and 187 non-argument tweets for Brexit (the test set). Below we report an example of argument tweet (a), and of a non-argument tweet (b).

(a) *Junker asks "who does he think I am". I suspect elected PM Tsipras thinks Junker is an unelected Eurocrat. #justsaying #democracy #grexit*
(b) *#USAvJPN #independenceday #JustinBieberBestIdol Macri #ConEsteFrioYo happy 4th of july #Grefenderum Wireless Festival*

We cast the argument detection task as a binary classification task, and we apply the supervised algorithms described in Section 5.5. Table 5.15 reports on the obtained results with the different configurations, while Table 5.16 reports on the results obtained by the best configuration, i.e., LR + All features, per each category.

Approach	Precision	Recall	F1
RF+L	0.76	0.69	0.71
LR+L	0.76	0.71	0.73
LR+all features	0.80	0.77	**0.78**

Table 5.15: Results obtained on the test set for the argument detection task (L=lexical features)

Category	P	R	F1	#arguments per category
non-arg	0.46	0.60	0.52	187
arg	0.89	0.82	0.85	713
avg/total	0.80	0.77	**0.78**	900

Table 5.16: Results obtained by the best model on each category of the test set for the argument detection task

Most of the miss-classified tweets are either ironical, e.g.:

If #Greece had a euro for every time someone mentioned #Grexit and #Greferendum they would probably have enough for a bailout. #GreekCrisis

that was wrongly classified as argument, or contain reported speech, e.g.:

Jeremy Warner: Unintentionally, the Greeks have done themselves a favour. Soon, they will be out of the euro `http://t.co/YmqXi36lGj` *#Grexit*

that was wrongly classified as non argument. Our results are comparable to those reported in [59] (they trained a supervised classifier on the tweets of all topics in the DART dataset but the iWatch, used as test

set). Better performances obtained in our setting are most likely due to a better feature selection, and to the fact that in our case the topics in the training and test sets are more homogeneous.

Task 2: Factual vs opinion classification

This task consists in classifying argument-tweets as containing factual information or being opinion-based [251]. Our interest focuses in particular on factual argument-tweets, as we are interested then in the automated identification of their sources. This would allow then to rank factual tweet-arguments depending on the reliability or expertise of their source for subsequent tasks as fact checking. Given the huge amount of work in the literature devoted to opinion extraction, we do not address any further analysis on opinion-based arguments here, referring the interested reader to [208].

An argument is annotated as *factual* if it contains a piece of information which can be proved to be true (see example (a) below), or if it contains "reported speech" (see example (b) below). All the other argument tweets are considered as "opinion" (see example (c) below).

(a) *72% of people who identified as "English" supported #Brexit (while no majority among those identifying as "British")* `https://t.co/MuUXqncUBe`
(b) *#Hollande urges #UK to start #Brexit talks as soon as possible.* `https://t.co/d12TV8JqYD`.
(c) *Trump is going to sell us back to England. #Brexit #RNCinCLE*

Our dataset contains 230 factual argument tweets and 516 opinion argument tweets for Grexit (training set), and 138 factual argument tweets and 575 opinion argument tweets for Brexit (test set).

To address the task of factual vs opinion arguments classification, we apply the supervised classification algorithms described in Section 5.5. Tweets from Grexit dataset are used as training set, and those from Brexit dataset as test set. Table 5.17 reports on the obtained results, while Table 5.18 reports on the results obtained by the best configuration, i.e. LR + All features, per each category.

Approach	Precision	Recall	F1
RF+L	0.75	0.68	0.71
LR+L	0.75	0.75	0.75
LR+all features	0.81	0.79	**0.80**

Table 5.17: Results obtained on the test set for the factual vs opinion argument classification task (L=lexical features)

Most of the miss-classified tweets contain reported opinions/reported speech and are wrongly classified by the algorithm as opinion - such behaviour could be expected given that sentiment features play a major role in these cases, e.g.,

Thomas Piketty accuses Germany of forgetting history as it lectures Greece `http://t.co/B0UqPn0i6T` *#grexit*

Again, the other main reason for miss-classification is sarcasm/irony contained in the tweets, e.g.,

Category	P	R	F1	#arguments per category
fact	0.49	0.50	0.50	138
opinion	0.88	0.87	0.88	575
avg/total	0.81	0.79	**0.80**	713

Table 5.18: Results obtained by the best model on each category of the test set for the factual vs opinion argument classification task

So for Tsipras, no vote means back to the table, for Varoufakis, meant Grexit?

that was wrongly classified as fact.

Task 3: Source identification

Since factual arguments (as defined above) are generally reported by news agencies and individuals, the third task we address - and that can be of a value in the context of social media - is the recognition of the information source that disseminates the news reported in a tweet (when explicitly mentioned). For instance, in:

The Guardian: Greek crisis: European leaders scramble for response to referendum no vote. `http://t.co/cUNiyLGfg3`

the source of information is The Guardian newspaper. Such annotation is useful to rank factual tweet-arguments depending on the reliability or expertise of their source in news summarization or fact-checking applications, for example.

Our dataset contains 79 factual argument tweets where the source is explicitly cited for Grexit (training set), and 40 factual argument tweets where the source is explicitly cited for Brexit (test set). Given the small size of the available annotated dataset, to address this task we implemented a simple string matching algorithm that relies on a gazetteer containing a set of Twitter usernames and hashtags extracted from the training data, and a list of very common news agencies (e.g. BBC, CNN, CNBC). If no matches are found, the algorithm extracts the NEs from the tweets through [236]'s system, and applies the following two heuristics: *i)* if a NE is of type `dbo:Organisation` or `dbo:Person`, it considers such NE as the source; *ii)* it searches in the abstract of the DBpedia[50] page linked to that NE if the words "news", "newspaper" or "magazine" appear (if found, such entity is considered as the source). In the example above, the following NEs have been detected in the tweet: "The Guardian" (linked to the DBpedia resource `http://dbpedia.org/page/The_Guardian`) and "Greek crisis" (linked to `http://dbpedia.org/page/Greek_government-debt_crisis`). Applying the mentioned heuristics, the first NE is considered as the source. Table 5.19 reports on the obtained results. As baseline, we use a method that considers all the NEs detected in the tweet as sources.

[50] `http://www.dbpedia.org`

Approach	Precision	Recall	F1
Baseline	0.26	0.48	0.33
Matching+heurist.	0.69	0.64	**0.67**

Table 5.19: Results obtained on the test set for the source identification task

Most of the errors of the algorithm are due to information sources not recognized as NEs (in particular, when the source is a Twitter user), or NEs that are linked to the wrong DBpedia page. However, in order to draw more interesting conclusions on the most suitable methods to address this task, we would need the increase the size of the dataset.

5.6 Argument mining on political speeches

In recent years, the analysis of argumentation using Natural Language Processing methods, so-called *argument mining* [158], has gained a lot of attention in the Artificial Intelligence research community and has been applied to a number of domains, from student essays [290] to scientific articles [299] and online user-generated content [314, 165]. However, while some of these approaches have been proposed to detect claims in political debates, e.g. [205, 233], little attention has been devoted to the prediction of relations between arguments, which could help historians, social and political scientists in the analysis of argumentative dynamics (e.g., supports, attacks) between parties and political opponents. For example, this analysis could support the study of past political speeches and of the repercussions of such claims over time. It could also be used to establish relations with the current way of debating in politics. In order to find argumentation patterns in political speeches, typically covering a wide range of issues from international politics to environmental challenges, the application of computational methods to assist scholars in their qualitative analysis is advisable.
In this work, we tackle the following research question: *To what extent can we apply argument mining models to support and ease the analysis and modeling of past political speeches?* This research question breaks down into the following subquestions:

- Given a transcription of speeches from different politicians on a certain topic, how can we automatically predict the relation holding between two arguments, even if they belong to different speeches?
- How can the output of the above-mentioned automated task be used to support history and political science scholars in the curation, analysis and editing of such corpora?

This issue is investigated by creating and analysing a new annotated corpus for this task, based on the transcription of discourses and official declarations issued by Richard Nixon and John F. Kennedy during the 1960 US Presidential campaign. Moreover, we develop a relation classification system with specific features able to *predict support* and *attack* relations between arguments [206], distinguishing them from unrelated ones. This argumentation mining pipeline ends with the visualization of the resulting graph of the debated topic using the OVA$^+$ tool.[51]

The main contributions of this section are *(1)* an annotated corpus consisting of 1,462 pairs of arguments in natural language (around 550,000 tokens) covering 5 topics, *(2)* a feature-rich Support Vector Machines

[51] http://ova.arg-tech.org/

(SVM) model for relation prediction, and *(3)* an end-to-end workflow to analyse arguments that, starting from one or more monological corpora in raw text, outputs the argumentation graph of user-defined topics.

To the best of our knowledge, there are no approaches in the argument mining literature that tackle the problem of relation prediction over political speeches. The most important feature of such speeches is their monological nature, with unaligned arguments, while debates are typically characterised by two interlocutors answering each other. This leads to more implicit attack and support relations between the arguments put forward by the candidates.[52] Applying the argument mining pipeline, and more precisely, the relation prediction stage to such speeches is the goal of our contribution.

Corpus Extraction and Annotation

Since no data for this task are available, we collect the transcription of speeches and official declarations issued by Nixon and Kennedy during 1960 Presidential campaign from The American Presidency Project.[53] The corpus includes 881 documents, released under the NARA public domain license, and more than 1,6 million tokens (around 830,000 tokens for Nixon and 815,000 tokens for Kennedy). We select this document collection because of its relevance from a historical perspective: the 1960 electoral campaign has been widely studied by historians and political scientists, being the first campaign broadcast on television. The issues raised during the campaign shaped the political scenario of the next decades, for example the rising Cold War tensions between the United States and the Soviet Union or the relationship with Cuba.

Dataset creation. In order to include relevant topics in the dataset, we asked a history scholar to list a number of issues that were debated during 1960 campaign, around which argumentation pairs could emerge. With his help, we selected the following ones: *Cuba*, *disarmament*, *healthcare*, *minimum wage* and *unemployment* (henceforth *topics*). We then extracted pairs of candidate arguments as follows. For each topic, we manually define a set of keywords (e.g., *[medical care, health care]*) that lexically express the topic. Then, we extract from the corpus all sentences containing at least one of these keywords, plus the sentence before and after them to provide some context: each candidate argument consists then of a snippet of text containing three consecutive sentences and a date, corresponding to the day in which the original speech was given during the campaign.

In the following step, we combine the extracted snippets into pairs using two different approaches. Indeed, we want to analyse two different types of argumentations: those *between candidates*, and those emerging from the speeches uttered by the *same candidate* over time. In the first case, for each topic, we sort all the candidate arguments in chronological order, and then create pairs by taking one or more snippets by a politician and the one(s) immediately preceding it by his opponent. These data are thus shaped as a sort of indirect dialogue, in which Nixon and Kennedy talk about the same topics in chronological order. However, the arguments of a speaker are not necessarily the direct answer to the arguments of the other one, making it challenging to label the relation holding between the two.

In the second case, we sort by topic all the candidate arguments in chronological order, as in the previous approach. However, each candidate argument is paired with what the same politician said on the same topic in the immediately preceding date. These data provide information about how the ideas of Nixon and Kennedy evolve during the electoral campaign, showing, if any, shifts in their opinions. We follow these two approaches also with the goal to obtain a possibly balanced dataset: we expect to have more attack

[52] In argument mining, a support is a statement (source of the relation) that underpins another statement (target of the relation). It holds between a target and a source statement if the source statement is a justification or a reason for the target statement.

[53] The American Presidency Project (`http://www.presidency.ucsb.edu/1960_election.php`)

relations holding between pairs of arguments from different candidates, while pairs of arguments from the same candidate should be coherent, mainly supporting each other.

Through this pairing process, we obtain 4,229 pairs for the *Cuba* topic, 2,508 pairs for *disarmament*, 3,945 pairs for *health-care*, 6,341 pairs for *minimum wage*, and 2,865 pairs for *unemployment*, for a total of 19,888 pairs.

Annotation. From the pool of automatically extracted pairs, we manually annotate a subset of 1,907 pairs randomly selected over the five topics. Annotators were asked to mark if between two given arguments there was a relation of *attack* (see Example 42 on minimum wage), a relation of *support* (see Example 43 on disarmament) or if there was *no relation* (arguments are neither supporting, nor attacking each other, tackling different issues of the same topic).

Example 42.
Nixon: And here you get the basic economic principles. If you raise the minimum wage, in my opinion - and all the experts confirm this that I have talked to in the Government - above $1.15, it would mean unemployment; unemployment, because there are many industries that could not pay more than $1.15 without cutting down their work force. $1.15 can be absorbed, and then at a later time we could move to $1.25 as the economy moves up.
Kennedy: The fact of the matter is that Mr. Nixon leads a party which has opposed progress for 25 years, and he is a representative of it. He leads a party which in 1935 voted 90 percent against a 25-cent minimum wage. He leads a party which voted 90 percent in 1960 against $1.25 an hour minimum wage.

Example 43.
Nixon: I want to explain that in terms of examples today because it seems to me there has been a great lack of understanding in recent months, and, for that matter in recent years, as to why the United States has followed the line that it has diplomatically. People have often spoken to me and they have said, Why can't we be more flexible in our dealings on disarmament? Why can't we find a bold new program in this area which will make it possible for the Soviet Union to agree? The answer is that the reason the Soviet Union has not agreed is that they do not want apparently to disarm unless we give up the right to inspection.
Nixon: People say, Now, why is it we can't get some imaginative disarmament proposals, or suspension of nuclear test proposals? Aren't we being too rigid? And I can only say I have seen these proposals over the years, and the United States could not have been more tolerant. We have not only gone an extra mile - we have gone an extra 5 miles - on the tests, on disarmament, but on everything else, but every time we come to a blocking point, the blocking point is no inspection, no inspection.

The annotation guidelines included few basic instructions: if the statements cover more than one topic, annotators were asked to focus only on the text segments dealing with the chosen topic. Annotation was carried out by strictly relying on the content of the statements, avoiding personal interpretation. Examples of *attack* are pairs where the candidates propose two different approaches to reach the same goal, where they express different considerations on the current situation with respect to a problem, or where they have a different attitude with respect to the work done in the past. For example, in order to increase minimum wage, Nixon proposed to set it to 1.10$ per hour, while Kennedy opposed this initiative, claiming that 1.35$ should be the minimum wage amount. In this example, the opponents have the same goal, i.e., increase minimum wage, but their statements are annotated as an attack because their initiatives are different, clearly expressing their disagreement.

After an initial training following the above guidelines, 3 annotators were asked to judge a common subset of 100 pairs to evaluate inter-annotator agreement. This was found to be 0.63 (Fleiss' Kappa), which as a rule of thumb is considered a substantial agreement [192]. After that, each annotator judged a different set of argument pairs, with a total of 1,907 judgements collected. In order to balance the data, we discarded part of the pairs annotated with *no relation* (randomly picked).

Overall, the final annotated corpus[54] is composed of 1,462 pairs: 378 pairs annotated with *attack*, 353 pairs annotated with *support*, and 731 pairs where these relations do not hold. An overview of the annotated corpus is presented in Table 5.20.

Topic	Attack	Support	No Relation
Cuba	38	40	180
Disarmament	76	108	132
Medical care	75	72	142
Minimum wage	125	80	107
Unemployment	64	53	170

Table 5.20: Topic and class distribution in the annotated corpus

Experiments on Relation Prediction

To facilitate the construction of argument graphs and support the argumentative analysis of political speeches, we propose an approach to automatically label pairs of arguments according to the relation existing between them, namely *support* and *attack*.

Given the strategy adopted to create the pairs, the paired arguments may happen to be also unrelated (50% of the pairs are labeled with *no relation*). Therefore, we first isolate the pairs connected through a relation, and then we classify them as *support* or *attack*. Each step is performed by a binary classifier using specific features, which we describe in the following subsection. In the section, we present the results obtained with the feature set that achieved the best performance on 10-fold cross validation.

Experimental setting. The *first step* concerns the binary classification of related and unrelated pairs. In this step the pairs annotated with support and attack have been merged under the *related* label. We first pre-process all the pairs using the Stanford CoreNLP suite [214] for tokenization, lemmatization and part-of-speech tagging. Then, for each pair we define three sets of features, representing the lexical overlap between snippets, the position of the topic mention in the snippet, as a proxy for its relevance, and the similarity of snippets with other related / unrelated pairs.

Lexical overlap: the rationale behind this information is that two related arguments are supposed to be more lexically similar than unrelated ones. Therefore, we compute *i)* the number of nouns, verbs and adjectives shared by two snippets in a pair, normalized by their length, and *ii)* the normalized number of nouns, verbs and adjectives shared by the argument subtrees where the topic is mentioned.

Topic position: the rationale behind this information is that, if the same topic is central in both candidate arguments, then it is likely that these arguments are related. To measure this, we represent with a set of

[54]The dataset is available at `https://dh.fbk.eu/resources/political-argumentation`

features how often the topic (expressed by a list of keywords, see previous section on dataset creation) appears at the beginning, in the central part or at the end of each candidate argument.

Similarity with other related / unrelated pairs: the intuition behind this set of features is that related pairs should be more similar to other related pairs than to unrelated ones. For each topic, its merged *related* and *unrelated* pairs are represented as two vectors using a bag-of-words model. Their semantic similarity with the individual pairs in the dataset is computed through cosine similarity and used as a feature.

For classification, we adopt a supervised machine learning approach training Support Vector Machines with radial kernel using LIBSVM [100].

In the *second step* of the classification pipeline, we take in input the outcome of the first step and classify all the pairs of related arguments as support or attack. We rely on a set of surface, sentiment and semantic features inspired by Menini and Tonelli (2016) and Menini et al. (2017). We adopt the **Lexical overlap** set of features used also for the first step, to which we add the features described below. In general, we aim at representing more semantic information compared to the previous step, in which lexical features were already quite informative.

Negation: this set of features includes the normalized number of words under the scope of a negation in each argument, and the percentage of overlapping lemmas in the negated phrases of the two arguments.

Keyword embeddings: we use word2vec [225] to extract from each argument a vector representing the keywords of a topic. These vectors are extracted using the continuous bag-of-word algorithm, a windows size of 8 and a vector dimensionality of 50.

Argument entailment: these features indicate if the first argument entails the second one, and vice-versa. To detect the presence of entailment we use the Excitement Open Platform [213].

Argument sentiment: a set of features based on the sentiment analysis module of the Stanford CoreNLP suite [288] are used to represent the sentiment of each argument, calculated as the average sentiment score of the sentences composing it.

Additional features for lexical overlap, entailment and sentiment are obtained also considering only the subtrees containing a topic keyword instead of the full arguments. The feature vectors are then used to train a SVM with radial kernel with LIBSVM, like in the first classification step.

Evaluation. We test the performance of the classification pipeline using the 1,462 manually annotated pairs with 10-fold cross-validation. The first classification step separates the argument pairs linked by either an *attack* or a *support* relation from the argument pairs with *no relation* (that will be subsequently discarded). The purpose of this first step is to pass the related pairs to the *second step*. Thus, we aim at the highest precision, in order to minimise the number of errors propagated to the second step. Table 5.21 shows the results of the classification for the first step. We choose a configuration that, despite a low recall (0.23), scores a precision of 0.88 on the *attack/support* pairs, providing for the second step a total of 194 argument pairs.

	Unrelated	Attack/Support	Average
Precision	0.56	0.88	0.72
Recall	0.97	0.23	0.60
F1	0.71	0.36	0.65

Table 5.21: Step 1: classification of related / unrelated pairs

The second step classifies the related pairs assigning an *attack* or a *support* label. We provide two evaluations: we report the classifier performance only on the gold *attack* and *support* pairs (Table 5.22), and on the pairs classified as related in the first step (Table 5.23). In this way, we evaluate the classifier also in a real setting, to assess the performance of the end-to-end pipeline.

	Attack	Support	Average
Precision	0.89	0.75	0.82
Recall	0.79	0.86	0.83
F1	0.84	0.80	0.82

Table 5.22: Step 2: classification of *Attack* and *Support* using only gold data.

	Attack	Support	Average
Precision	0.76	0.67	0.72
Recall	0.79	0.86	0.83
F1	0.77	0.75	0.77

Table 5.23: Step 2: classification of *Attack* and *Support* using the output of Step 1.

As expected, accuracy using only gold data is 0.82 (against a random baseline of 0.70), while it drops to 0.72 (against a random baseline of 0.51) in the real-world setting. We also test a 3-class classifier, with the same set of features used in the two classification steps, obtaining a precision of 0.57. This shows that *support/attack* and *no relation* are better represented by using different sets of features, therefore we opt for two binary classifiers in cascade.

Notice that a comparison of our results with existing approaches to predict argument relations, namely the approach of [291] on persuasive essays, cannot be fairly addressed due to huge differences in the complexity of the used corpus. With their better configuration, [291] obtain an F1 of 0.75 on persuasive essays (that are a very specific kind of texts, human upperbound: macro F1 score of 0.854), and of 0.72 on microtexts [256]. The difference in the task complexity is highlighted also in the inter-annotator agreement. Differently from persuasive essays, where students are requested to put forward arguments in favour and against their viewpoint, in political speeches, candidates often respond to opponents in subtle or implicit ways, avoiding a clear identification of opposing viewpoints.

Error analysis. If we analyse the classifier output at topic level, we observe that overall the performance is consistent across all topics, with the exception of *minimum wage*. In this latter case, the classifier performs much better, with an accuracy of 0.94 in the second step. This is probably due to the fact that Kennedy's and Nixon's statements about minimum wage are very different and the discussion revolves around very concrete items (e.g., the amounts of the minimum wage, the categories that should benefit from it). In other cases, for example disarmament or Cuba, the speakers' wording is very similar and tends to deal with abstract concepts such as freedom, war, peace.

Furthermore, we observe that the classifier yields a better performance with argument pairs by the same person rather than those uttered by different speakers: in the first case, accuracy is 0.86, while in the second one it is 0.79 (Step 2).

Looking at misclassified pairs, we notice very challenging cases, where the presence of linguistic devices like rhetorical questions and repeated negations cannot be correctly captured by our features. Example 44 reports on a pair wrongly classified as *Support* belonging to the *health care* topic:

Example 44.
Nixon: *Now, some people might say, Mr. Nixon, won't it be easier just to have the Federal Government take this thing over rather than to have a Federal-State program? Won't it be easier not to bother with private health insurance programs? Yes; it would be a lot simpler, but, my friends, you would destroy the standard of medical care.*
Kennedy: *I don't believe that the American people are going to give their endorsement to the leadership which believes that medical care for our older citizens, financed under social security, is extreme, and I quote Mr. Nixon accurately.*

Visualization and Analysis of the Argumentation Graphs

In this section, we describe how the results of our relation prediction system are then used to construct the argumentation graphs about the debated topics.

Several tools have been proposed to visualize (and then reason upon) argumentation frameworks in the computational argumentation field, e.g., Carneades[55], GRAFIX[56], and ConArg2[57]. However, two main problems arise when trying to use such tools for our purposes: first, they are not tailored to long, natural language snippets (the usual names of arguments in computational argumentation are of the form arg_1), and second, they do not consider the possibility to identify specific argumentation schemes over the provided text. For all these reasons, we decided to rely upon a well-know tool called OVA$^+$ [178], an on-line interface for the manual analysis of natural language arguments. OVA$^+$ grounds its visualization on the Argument Interchange Format (AIF) [105], allowing for the representation of arguments and the possibility to exchange, share and reuse the resulting argument maps. OVA$^+$ handles texts of any type and any length.

The last step of our argument mining pipeline takes in input the labeled pairs returned by the relation prediction module and translates this output to comply with the AIF format. This translation is performed through a script converting the `CSV` input file into `json` file to be load on OVA$^+$ through its online interface.[58] In this mapping, each argument is extracted in order to create an information node (I-node) [105], and then, it is possible to create the associated locution node (L-node) and to specify the name of the speaker. The locution appears, preceded by the name of the participant assigned to it, and edges link the L-node to the I-node via an "Asserting" YA-node, i.e., the illocutionary forces of locutions, as in the Inference Anchoring Theory (IAT) model [64]. Supports or attacks between arguments are represented as follows, always relying upon the standard AIF model. A RA-node (*relation of inference*) should connect two I-nodes. To elicit an attack between two arguments, RA-nodes are changed into CA-nodes, namely *schemes of conflict*. Nodes representing the support and the attack relations are the "Default Inference" and the "Default Conflict"

[55] `http://carneades.github.io/`
[56] `https://www.irit.fr/grafix`
[57] `http://www.dmi.unipg.it/conarg/`
[58] The script and the argumentation graphs about the five topics in our corpus (both gold standard and system's output) are available at `https://dh.fbk.eu/resources/political-argumentation`

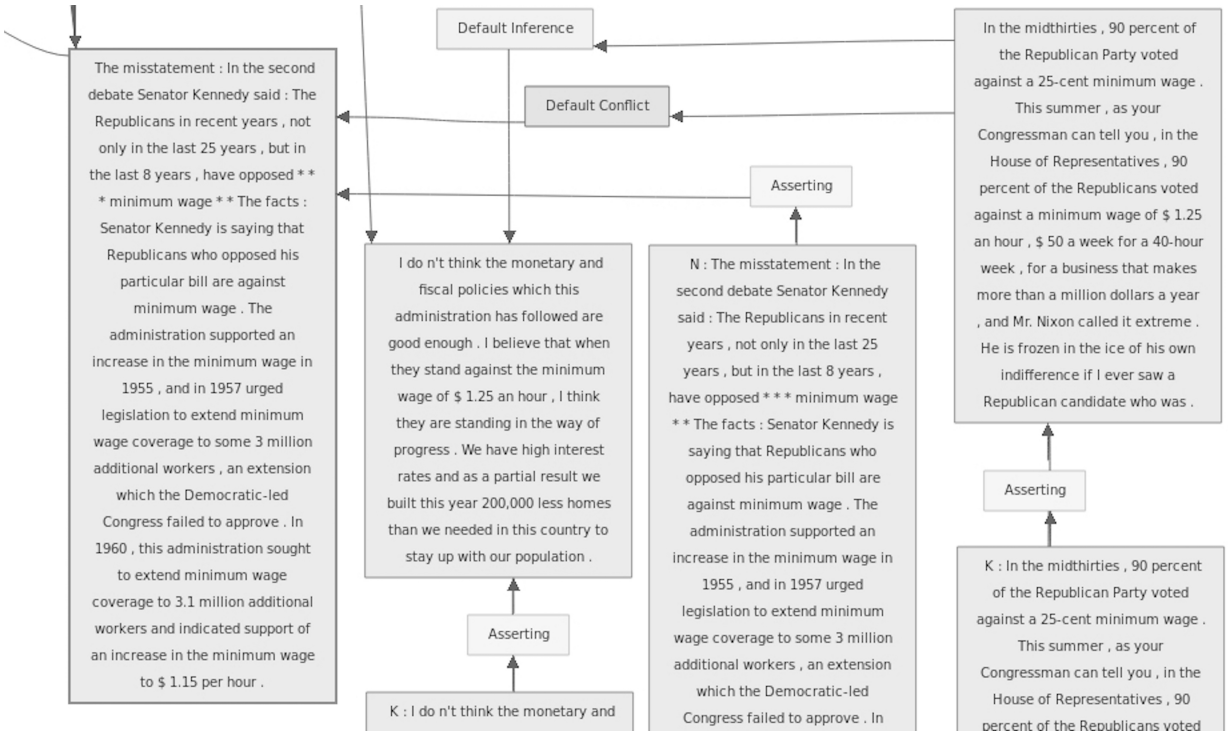


Figure 5.12: The argumentation graph about the topic *minimum wage* visualized through the OVA$^+$ tool.

nodes, respectively. Figure 5.12 shows (a portion of) the argumentation graph resulting from the relation prediction step about the topic *minimum wage*, where three I-nodes (i.e., arguments) are involved in one support and one attack relation. The *Asserting* nodes connect each argument with its own source (e.g., K for Kennedy and N for Nixon).

OVA$^+$ allows users to load an analysis, and to visualize it. Given the loaded argumentation graph, the user is supported in analyzing the graph by identifying argumentation schemes [318], and adding further illocutionary forces and relations between the arguments. This final step substantially eases the analysis process by historians and social scientists. Moreover, at the end of the analysis, OVA$^+$ permits to save the final argumentation graph on the user's machine (image or json file).

This graph-based visualization is employed to support political scientists and historians in analysing and modeling political speeches. This proves the usefulness of applying the argumentation mining pipeline over such kind of data: it allows users to automatically identify, among the huge amount of assertions put forward by the candidates in their speeches, the main points on which the candidates disagree (mainly corresponding to the solutions they propose to carry out or their own viewpoints on the previous administrations' effectiveness) or agree (mainly, general-purpose assertions about the country's values to promote).

In the following, we analyze the argumentative structure and content of two of the graphs resulting from the discussed topics (i.e., *minimun wage* and *health care*), highlighting main conflicting arguments among candidates, and other argumentative patterns. Note that this analysis is carried out on the proposed dataset, that contains a subset of all the speeches of the candidates, but gives a clear idea of the kind of analysis that could be performed by scholars on the entirety of the speeches. In general (and this is valid for all the

analyzed graphs), we notice that the candidates almost always disagree either on the premises (e.g., who caused the problem to be faced) or on the proposed solutions (the minor claims).

Minimum wage. A widely discussed topic by both candidates was *minimum wage*, i.e., the bill to set the lowest remuneration that employers may legally pay to workers. It is worth noticing that the argumentation graph for the minimum wage corpus is rather complicated, and it highlights some main controversial issues. The candidates do not agree about the causes of the low minimum wage in 1960 in the US. More precisely, Kennedy attacks the fact that the administration supported an increase in the minimum wage by attacking Nixon's argument *"The misstatement: In the second debate Senator Kennedy said: The Republicans in recent years, not only in the last 25 years, but in the last 8 years, have opposed minimum wage. The facts: [. . .] The administration supported an increase in the minimum wage in 1955, and in 1957 urged legislation to extend minimum wage coverage to some 3 million additional workers, an extension which the Democratic-led Congress failed to approve. In 1960, this administration sought to extend minimum wage coverage to 3.1 million additional workers and indicated support of an increase in the minimum wage to $1.15 per hour."*. This argument is attacked from different perspectives, leading to a disagreement on the actions the administration carried out in the past years to deal with the minimum wage problem. For instance, as shown in Figure 5.12, Kennedy states that *"In the midthirties, 90 percent of the Republican Party voted against a 25-cent minimum wage. This summer, as your Congressman can tell you, in the House of Representatives, 90 percent of the Republicans voted against a minimum wage of $1.25 an hour, $50 a week for a 40-hour week, for a business that makes more than a million dollars a year, and Mr. Nixon called it extreme. He is frozen in the ice of his own indifference"*. While we may say that this source of disagreement is about the causes of the minimum wage issue, another main source of disagreement is represented by the solutions proposed by the two candidates, which mainly differ regarding the amount of increase of the minimum wage and the coverage of the two respective bills. All these issues become evident with ease in the resulting argumentation graph about the minimum wage topic.

Medical care. The problem of medical care for the elderly was a main problem in 1960, and this topic was widely discussed in the campaign. The resulting argumentation graph highlights some relevant argumentative patterns that are worth analyzing. In general, in the argumentation graphs we are analyzing, the support relation holds between arguments proposed by the same candidate, ensuring in this way a certain degree of coherence in their own argumentation. Interestingly, in the argumentation graph on the topic *medical care*, we can observe that a support relation holds between an argument from Kennedy and one from Nixon, i.e., *"Those forced to rely on surplus food packages should receive a more balanced, nourishing diet. And to meet the pressing problem confronting men past working age, and their families, we must put through an effective program of medical care for the aged under the social security system. The present medical care program will not send one penny to needy persons without further action by the Congress and the State legislatures."* supports *"N: We stand for programs which will provide for increased and better medical care for our citizens, and particularly for those who need it, who are in the older age brackets - and I will discuss that more a little later. We stand for progress in all of these fields, and certainly, as I stand here before you, I am proud to be a part of that platform and of that program"*. These instances of support among candidates mostly concern general issues, i.e., a program of medical care for the elderly is needed.

In summary, our system allows the detection of such argumentation patterns (i.e., topics on which both candidates agree or disagree, topics on which they provide contradictory assertions) and the analysis of how they connect with the other statements asserted in the speeches. As for future work, we face two

major challenges. First, to improve the system performances, we need a finer-grained argument boundary definition. Namely, the goal is to identify within an argument its *evidences* and *claims*, so that the relations of support and attack may also be addressed towards these precise argument components. This would also have an impact on facilitating the work of scholars in the manual analysis of the argumentation graphs generated by our system. Second, we plan to evaluate the system with scholars in history and political sciences, who will be asked to judge not only the classification output but also the way in which it is displayed. We are currently working at a more interactive interface to display graphs with their textual content, so that users can select and visualize subgraphs according to the selected argumentative pattern.

5.7 Related work

Argument mining on online debate platforms

Among the set of online debate systems, Debategraph[59] is an online system for debates supporting the incremental development of argument structures, but it is not grounded on argument theory to decide the accepted arguments.

Gilbert [154] addresses the topic of human/computer argumentation, where the ability to identify and classify various locutions as facts, values and goals is discussed. The paper does not present a solution to the problem of automatically generating the arguments from NL text. The author grounds his observations on Toulmin [302] argumentation model.

Chesnevar et al. [104] use defeasible argumentation to assist the language usage assessment. Their system provides recommendations on language patterns using indices (computed from Web corpora) and defeasible argumentation, where the preference criteria for language usage are formalized as defeasible and strict argumentation rules. The aim of the paper is different from ours. No NL techniques are used to automatically detect and generate the arguments.

Carenini et al. [85] present a computational framework for generating evaluative arguments. The framework, based on the user's preferences, produces the arguments following the guidelines of argumentation theory to structure and select evaluative arguments. Then, a natural language processing step returns the argument in natural language. The output of the argumentation strategy is a text plan indicating the propositions to include in the argument and its overall structure. The aim of the paper is different from our one: we do not use natural language generation to produce the arguments, but we use textual entailment to detect the arguments in natural language text. We use the word "generation" with the meaning of generation of the abstract arguments from the text, and not with the meaning of natural language generation. Concerning argumentation, we use bipolar argumentation to reason over the arguments to identify the accepted ones. We do not address argumentation-based persuasion or planning.

Leite et al. [197] envision a self-managing online debating system able to accommodate different kinds of participation of the agents. However, while we re-use Dung's abstract theory, they depart from this approach and defend the view that these debates should provide more than an accepted/rejected classification of the issue at stake. They do not apply natural language processing techniques to identify the arguments in natural language debates.

Wyner et al. [325] present a policy making support tool based on forums. They propose to couple NLP and argumentation to provide the set of well structured statements that underlie a policy. Apart from the different goal of this work, there are several points which distinguish our proposal from this one. First,

[59] `http://debategraph.org`

their NLP module guides the participant in writing the input text using Attempt to Controlled English which allows the usage of a restricted grammar and vocabulary. After parsing the text, the sentences are translated to FOL. We do not have any kind of lexicon or grammar restriction, and we do not support the participant in writing the text, but we automatically extract the arguments from the debates. Second, the inserted statements are associated with a mode indicating the relation between the existing statements and the input statement. We do not ask the participants to explicit the relation among the arguments, we infer them using TE. Moreover, no evaluation of their framework is provided.

Heras et al. [167] show how to model the opinions put forward on business oriented websites using argumentation schemes. The idea is to associate a scheme to each argument to have a formal structure which makes the reasoning explicit. We share the same goal, that is providing a formal structure to online dialogues to evaluate them, but, differently from [167], in our proposal we achieve this issue using an automatic technique to generate the arguments from natural language texts as well as their relations.

Rahwan et al. [268] present Avicenna, a Web-based system used to reason about arguments, ranging from automatic argument classification to reason about chained argument structures. In Avicenna, the arguments are inserted by participants through a form, and the participants can decide to attack or support existing arguments, while in our framework participants do not enter arguments: it automatically returns the abstract arguments, the relationships among them highlighting the accepted arguments.

Mochales et al. [231] experiment ML approaches to recognize features characterizing legal arguments. We adopt a more general framework, i.e. TE (implementable also using ML techniques) to extract open-domain arguments, and automatically assign their relations.

Argument mining on Wikipedia history

A few works investigate the use of Wikipedia revisions in NLP tasks. In Zanzotto and Pennacchiotti [330], two versions of Wikipedia and semi-supervised machine learning methods are used to extract large TE data sets, while Cabrio et al. [66] propose a methodology for the automatic acquisition of large scale context-rich entailment rules from Wikipedia revisions. [328] focus on using edit histories in Simple English Wikipedia to extract lexical simplifications. Nelken and Yamangil [326] compare different versions of the same document to collect users' editorial choices, for automated text correction and text summarization systems. Max and Wisniewski [219] create a corpus of natural rewritings (e.g. spelling corrections, reformulations) from French Wikipedia revisions. Dutrey et al. [138] analyze part of this corpus to define a typology of local modifications.

Other approaches couple NLP and argumentation. Chasnevar and Maguitman [104] use defeasible argumentation to assist the language usage assessment. Their system provides recommendations on language patterns and defeasible argumentation. No natural language techniques are applied to automatically detect and generate the arguments. Carenini and Moore [85] present a complete computational framework for generating evaluative arguments. The framework, based on the user's preferences, produces the arguments following the guidelines of argumentation theory to structure and select evaluative arguments. Differently from their work, we do not use natural language generation to produce the arguments, but we use TE to detect the arguments in natural language text. We use the word "generation" with the meaning of generation of the abstract arguments from the text, and not with the meaning of NL generation. Wyner and van Engers [325] present a policy making support tool based on forums. They propose to couple NLP and argumentation to provide the set of well structured statements that underlie a policy. Beside the goals, several points distinguish the two works: *i)* their NLP module guides the user in writing the text using a restricted grammar and vocabulary, while we have no lexicon or grammar restrictions; *ii)* the inserted statements are

associated with a mode indicating the relation between the existing and the input statements. We do not ask the user to explicit the relation among the arguments, we infer them using TE; *iii)* no evaluation of their framework is provided. Heras et al. [167] show how to model the opinions on business oriented websites using argumentation schemes. We share the same goal (i.e. providing a formal structure to on-line dialogues for evaluation,), but in our proposal we achieve it using an automatic technique to generate the arguments from natural language texts as well as their relations.

Argument mining on Twitter

To the best of our knowledge, we are the first tackling the challenge of applying argument mining to Twitter data. Argumentation is applied to Twitter by Grosse et al. [161] who extract a particular version of arguments they called "opinions" based on incrementally generated queries. Their goal is to detect conflicting elements in an opinion tree to avoid potentially inconsistent information. Both the goal and the adopted methodology is different from the one we present in this chapter.

5.8 Conclusion

The aim of this chapter was to show how the joint effort of two, rather disjoint, research communities in AI resulted in the development of a new research area: *argument mining*. This synergy among researchers from both NLP and KRR communities has led to the conception and development of systems able to mine a variety of textual documents, e.g., legal cases, persuasive essays, online debates and tweets, to detect premises and claims, and predict the relations among them. The results obtained so far in AM have attracted the interest (and investment) of companies (e.g., IBM), and have raised high expectations for the future findings in the area.

Since the standard definition of the AM framework, in the last years, a number of new challenges have been proposed in the literature. In particular, Dusmanu et al. [136] select argumentative tweets and distinguish those conveying an opinion from those containing *factual* information, to detect their source of information (e.g., the BBC). Other colleagues [163, 258, 135] focused on arguments persuasion to study the relation "argument A is more convincing than argument B".

In addition, AM is strongly connected with hot topics in AI, as deep learning (heavily used in AM), fact checking and misinformation detection (the prediction of the attacks between arguments is a building block for fake news detection), and explanations of machine decisions (AM can disclose how the information on which the machine relies to make its own decisions is retrieved). Other scenarios where AM can contribute are medicine (where AM can detect information needed to reason upon clinical trials), politics (where AM can provide the means to automatically identify fallacies and unfair propaganda), and for cyberbullism prevention (where AM can support the detection of repeated attacks against a person[60]).

Alas, all that glitters is not gold, and some open issues in AM should be tackled to actually attain the expectations. First of all, system performances should improve. Despite the good results obtained in some application scenarios, i.e., persuasive essays [289] (where the structure of the essays themselves eases the argument component detection task), for other kinds of documents, e.g., legal cases [298] and microtexts [255], more work is still required. It is important to underline here that also human agreement (generally viewed as the upper bound on automatic performance in annotation tasks) is affected by the complexity of the AM tasks. Moreover, various heterogeneous datasets have been produced since the beginning

[60] http://creep-project.eu/

of research in AM. Because of the immaturity of a rising field, and the lack of clear definitions, each dataset has been annotated relying on slightly different definitions of argument components and of the relations holding between them, thus preventing the possibility of a straightforward alignment among datasets. While on the one side, it would be worth trying some kind of unification of existing resources, on the other side, this fact shows that the AM framework is flexible enough to adapt to different use case scenarios, e.g., premises and claims are not the same in legal cases, persuasive essays and Twitter. In [121], a qualitative analysis of six different datasets commonly used in AM is presented, to underline the different conceptualization of claims. The question about the worthiness of unifying the existing datasets is still open and under debate. [315] highlight and empirically study a related issue, i.e., the question of how different the theoretical (computational models of arguments) and practical views of argumentation quality actually are. Their results show that, on the one hand, most reasons for quality differences in practice seem well-represented in the theory, but on the other hand, some quality dimensions remain hard to assess in practice, resulting in a limited agreement.

Chapter 6

Emotions in human argumentation

6.1 Introduction

This chapter synthesizes my contributions in the area of argumentation and emotions, dealing with the analysis and study of how emotions impact on human argumentation and viceversa. These contributions are across the Cognitive Science research area and the computational models of argument one. These contributions have been published in several venues:

- *Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, Fabien Gandon. Emotions in Argumentation: an Empirical Evaluation. Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015): 156-163* [39],
- *Sahbi Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, Elena Cabrio. Persuasive Argumentation and Emotions: An Empirical Evaluation with Users. 19th International Conference on Human-Computer Interaction (HCI 2017): 659-671* [38],
- *Serena Villata, Sahbi Benlamine, Elena Cabrio, Claude Frasson, Fabien Gandon. Assessing Persuasion in Argumentation through Emotions and Mental States. Thirty-first International Florida Artificial Intelligence Research Society Conference (FLAIRS 2018)* [311],
- *Serena Villata, Elena Cabrio, Iméne Jraidi, Sahbi Benlamine, Maher Chaouachi, Claude Frasson, Fabien Gandon. Emotions and personality traits in argumentation: An empirical evaluation. Argument & Computation 8(1): 61-87 (2017)* [312].

The contributions presented in this chapter are the result of the collaboration with the Heron Laboratory at the University of Montreal, started in the context of the SEEMPAD project.

Understanding how humans reason and take decisions in debates and discussions is a key issue in cognitive science and a challenge for social applications. Moreover, with the growing importance of the Web, this issue is complicated by the fact that in such a hybrid space heterogeneous actors, both human and artificial, interact. As a typical example, Wikipedia is managed by users and bots who constantly contribute, agree, disagree, debate and update the content of the encyclopedia. In this context, several dimensions of the interaction affect the reasoning and decision making process, i.e., the arguments that are proposed online, the emotions of those who propose such arguments as well as the emotions of those reading these arguments, the social relationships among the involved actors, the writing of their messages, etc. This underlines the

need for multidisciplinary approaches and research for Web applications in general and for detecting and managing the emotional state of a user in particular to allow artificial and human actors to adapt their reactions to others' emotional states. It is also a useful indicator for community managers, moderators and editors to help them in handling the communities and the content they produce.

In this chapter, we argue that in order to apply argumentation to scenarios like e-democracy and online debate systems, designers must take both the argumentation and the emotions into account. In order to efficiently manage and interact with such a hybrid society, we need to improve our means to understand and link the different dimensions of the exchanges (e.g. social interactions, textual content of the messages, dialogical structures of the interactions, emotional states of the participants). Beyond the challenges individually raised by each dimension, a key problem is to link these dimensions and their analysis together with the aim to detect, for instance, a debate turning into a flame war, a content reaching an agreement, a good or bad emotion spreading in a community.

In this chapter, we aim to answer the research question: *What is the connection between the argumentation and the emotions in online debate interactions?* Such question breaks down into sub-questions:

- How are the arguments and their relations correlated with the polarity of the detected facial emotions?
- Is the relation between the kind and the amount of arguments proposed in a debate correlated with the mental engagement and workload detected for each participant in the debate?
- How do personality traits and opinions affect participants' emotions during the debate?

To answer these questions, we propose an empirical evaluation of the connection between argumentation, personality traits, and emotions in online debate interactions. This chapter describes an experiment with human participants which investigates the correspondences between the arguments and their relations put forward during a debate, the emotions detected by emotions recognition systems in the debaters, and the personality traits of the debaters. We designed an experiment where 12 debates were addressed by 4 participants each. Participants were asked to discuss about 12 topics in total proposed by moderators, e.g., "Religion does more harm than good" and "Cannabis should be legalized". Participants argue in plain English proposing arguments, that are in positive or negative relation with the arguments proposed by the other participants and by moderators. During these debates, participants are equipped with emotion detection devices, recording their emotions. Moreover, each participant filled in a questionnaire for Big Five personality traits [180]. We hypothesize that mental engagement and emotions are correlated to the argumentation holding in the debates, namely to the number of arguments that are proposed, and the kind of relations connecting them (i.e., support or attack). Moreover, we hypothesize that personality traits of debaters and debaters' opinions regarding the discussed topics modulate their emotional experiences during the debates.

In this chapter, we also present another study where we investigate how emotions and mental states impact on the persuasion strategies used by humans when they argue to each other. Also in this case, participants discuss on an online debate platform and they are equipped with emotion recognition and mental state detection tools.

A key point in our work is that, up to our knowledge, no user experiment has been carried out yet to determine what is the connection between the argumentation addressed during a debate, the emotions emerging in the participants, as well as their personality traits. An important result of the work reported here is the development of a publicly available dataset, capturing several debate situations, annotated with their argumentation structure and the emotional states automatically detected.

It is worth highlighting that bipolar abstract argumentation is used in our experimental setting to pair the arguments, connect them with the appropriate relation (either support or attack), and combine them in bipolar argumentation graphs. This structure allows for further reasoning activities over the data, where for instance the acceptability of the arguments depends on the mental engagement associated to their conception by their proposer, or a ranking is established over the acceptable arguments depending on the emotions (mostly positive, mostly negative, neutral) they generated in the audience. The definition and evaluation of these reasoning processes are not in the scope of the present chapter, and we left them for future research. It is worth clarifying also that, in this chapter, we are not interested in verifying explanation and reasoning theories proposed in cognitive psychology like, among others, those of Lombrozo and colleagues [323] about explanations in category learning, and Keil and colleagues [181] about explanatory reasoning through an abductive theory. We rely on bipolar argumentation for representing the debates, and to foster the application of reasoning techniques.

The chapter is organized as follows. In Section 6.2 we describe the main insights of the two components of our framework, then in Section 6.3, we describe the experimental protocol and the questionnaires we proposed to the debaters during the experiment, and our research hypotheses. In Section 6.4, we provide a detailed analysis of the experimental results. Section 6.5 discusses the field experiment we conducted to study the relation of emotions and mental states on argumentative persuasion strategies, and the obtained resuts. We compare this work with the relevant literature in Section 6.6. Conclusions end the chapter.

6.2 Emotions and personality traits in argumentation

In this section, we present the two main components involved in our experimental framework: *i) bipolar argumentation theory*, i.e., the formalism used to analyze and represent the argumentation elements from the debates, and *ii)* the methodologies and tools used to detect the degrees of attention, engagement, and workload of each participant involved in the debate, as well as her facial emotions.

Argumentation Theory

In order to analyze from the argumentation point of view the debates in which the participants to our experiment have been involved, we rely on abstract bipolar argumentation. In this way, we do not need distinguish the internal structure of the arguments (i.e., premises, conclusion), but we consider each argument proposed by the participants in the debate as a unique element, then analyzing the relation (positive or negative) it has with the other pieces of information put forward in the debate. The following example extracted from one of the debates addressed in our experiment shows how the arguments are connected to each other through a defeat or a support relation. Consider the following three arguments proposed by the participants of the debate about "Religion does more harm than good". We have that the issue of the debate is also our first argument whose proponent is the debate moderator, then the other two arguments are proposed by two different participants:

Argument 1 : *Religion does more harm than good.*

Argument 2 : *During all the existence of the human being, religion makes a lot of issues. It makes more hurts than cures.*

Argument 3 : *I think for people, religion is a refuge against the horrors of the world.*

Given such a kind of debate, we have that three arguments are proposed (namely Argument 1, Argument 2 and Argument 3) whose relations with each other are as follows: (Argument 2) supports (Argument 1), (Argument 3) attacks (Argument 1), and (Argument 3) attacks (Argument 2).

Note that in this chapter we are not interested in applying natural language processing techniques to detect automatically the relations among the arguments. On the contrary, we have manually built our data set of argumentation and emotions from the data retrieved during the debates of our experiment. Experiments with natural language processing approaches will be part of future work.

Emotion Detection

Human emotion analysis during traditional face-to-face or computer-mediated interaction has always been a challenging and attractive task mainly because of how emotions are closely associated to human behavior and experience. Several theories state that emotions serve as an adaptive function to our behavior, e.g., [148, 175, 194, 285]. Following these theories, the appraisal of an experience and the intention to act to maintain, adjust or change a condition related to this experience is impacted by emotions. During a debate, emotional reactions provoked by others' arguments could be, for example, a trigger for developing attacking or supporting arguments.

Emotion recognition methods can be categorized into three groups, each of them defining one level of how a usual emotional response is displayed, namely, *experiential*, *behavioral* and *physiological* [160]. For example if an individual is annoyed by someone else's argument, the subjective experience could be the anger; the behavioral response will be displayed through a higher voice tone (during a face-to-face conversation) or an angry facial expression; and the physiological response will be activated by an increasing level of heart rate. Usually, the experiential methods use subjective self-report instruments (such as surveys and questionnaires) to determine the emotion relative to a specific event. The behavioral methods are based on external observable clues detected from the individual's behavior (such as gestures, body movements, facial expression, voice tone and pitch, etc.) that can indicate the type of the emotion. The physiological methods rely on physical sensors (such as EDA, EEG, heart rate, respiration rate, temperature, etc.) to measure specific physiological shifts and patterns that can be related to specific emotions. In a computer-mediated context, the use of self-report surveys to recognize the individuals' emotions could be inconvenient. If these surveys are administrated at frequent intervals during the task, it could be disrupting for the task performance. However, non-frequent administration intervals of the survey could result into an undetailed and ambiguous assessment of the different emotions experienced during the task. Therefore, growing research interest has arisen towards using and combining behavioral and physiological methods for emotion recognition [265, 182, 183]. These methods allow automatic, objective and reasonably precise emotional recognition level.

In our study, we selected a behavioral method to extract the emotional manifestations. We used a set of webcams (one for each participant in the discussion) whose recordings have been analyzed with the FaceReader software[1] to detect a set of discrete emotions from facial expressions. Furthermore, we also recoded the EEG data from each participant in order to extract more complex information about their internal cognitive state. This cognitive information was aligned and analyzed jointly with the emotional information to have a global overview of the debate experience for each participant.

[1] `http://www.noldus.com/human-behavior-research/products/facereader`

Detecting emotions from facial expressions The emotional detection from facial expression is one of the most commonly and predominantly used methods [322, 8, 171]. In fact, facial expressions of basic emotions are widely believed to be naturally and universally expressed and recognized. In this study, we used the FaceReader software (version 6.0) to automatically extract the emotional reactions from the frame-by-frame videos recoded by the webcam. The FaceReader software launched by Noldus Information technology is able to recognize six basic emotions, namely, *happy*, *sad*, *angry*, *surprised*, *scared* and *disgusted* with an accuracy reaching 87%. The detection process is performed by extracting and classifying in real-time 500 key points in facial muscles of the target face. These key points are provided as input to a neural network trained on a dataset of 10000 manually annotated images corresponding to these six basic emotions. In addition to the probability of the presence of these six emotions, the software output also contains the probability of the neutral state as well as the *valence*, and the *arousal* of the emotional state. Information about the emotional valance defines the nature of the emotion and is ranging from -1 to $+1$. A positive valence value refers to pleasant emotion, whereas a negative valence value characterizes unpleasant emotions. The information about the arousal of the emotion defines its intensity and is also ranging from -1 to $+1$. A high arousal value indicates a high emotional intensity and a low value the opposite.

In this study, at each second in the debate, a dominant emotion is computed for each one of the four participants. This dominant emotion corresponds to the emotion (among the six detected by FaceReader) with the highest probability. Moreover, information about the valence and arousal of this emotion as well as their class (pleasant or unpleasant for the valence, high or low for the arousal) was also considered.

Emotiv EPOC EEG headset In order to record physiological data of the participants during the debate sessions, we used the 4 Emotiv Epoc EEG headsets (one for each participant). This device contains 14 electrodes spatially organized according to International $10-20$ system[2]. The pads of each electrode were moistened with a saline solution (contact lens cleaning solution) in order to enhance the quality of the signal. Figure 6.1 depicts the recorded sites, namely: AF3, AF4, F3, F4, F7, F8, FC5, FC6, P7, P8, T7, T8, O1, and O2. The reference of this EEG setup is represented by two other electrodes located behind the user's ears. The generated data are in (μV) with a 128Hz sampling rate. The signal frequencies are between 0.2 and 60Hz. An artifact rejection procedure based on the signal amplitude was performed in order to reduce the impact of blinking and body movement effect [146, 147]. More precisely, if the amplitude of any 1-s EEG in any site exceeds in 25% of the data point a predefined threshold, the segment is rejected.

Extracting the engagement index The term mental engagement refers to the level of attention and alertness during a task. The engagement index used in our study is based on the findings of [260] and [146]. In their study, it was found that the user's performance improved when an EEG index is used as a criterion for switching between manual and automated piloting mode. This index is computed from three EEG frequency bands: $\alpha(8-12hz)$, $\beta(12-22Hz)$ and $\theta(4-8hz)$:

$$eng = \frac{\theta}{\alpha+\beta}$$

.

This index is computed each second from the EEG signal. To smooth the values of this index and reduce its fluctuation, we used a moving average on a 40-second mobile window. More precisely, the value of the

[2]International $10-20$ system is an internationally recognized method to describe and apply the location of scalp electrodes in the context of an EEG test or experiment.

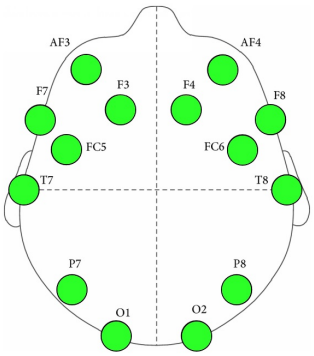


Figure 6.1: Emotiv Headset sensors/data channels placement.

index at time t corresponds to the total average of the ratios calculated on a period of 40 seconds preceding t. The extraction of the frequency bands (namely α, β and θ) was performed by multiplying every second of the EEG signal by a Hamming window (to reduce the spectral leakage) and applying a Fast Fourier Transform (FFT). As the Emotiv headset measures 14 regions at the same time, we used a combined value of α, β and θ frequency bands by summing their values over all the measured regions. To examine participants' engagement, we extract their minimum, average and maximum values during the debate, and we use such values to identify the range of engagement (High, Medium, Low) of every participant. Since its development by Pope and his colleagues, this engagement index has become a very important and popular technique for real time or offline tracking and analysis of individuals' engagement in several laboratory studies. In the educational sittings for example, this engagement index was used for monitoring learners' engagement and adapting learning activities according to their level of mental engagement by [294] and [102]. In robotics, this index was also used to leverage the interaction between a robot and a user by providing the robot real-time information about the user's engagement while the robot is speaking to him by [295]. The robot was successfully able to detect when the user is not anymore engaged in listening to him and tried to regain his attention by employing verbal and nonverbal techniques. This engagement index was also selected as a criterion for adapting a game's difficulty according to the player's level of engagement, showed promising results [99].

Extracting the workload index The term workload (or cognitive load in a learning context) refers to a measurable quantity of information processing demands placed on an individual by a task [249]. The mental workload is generally related to the working memory and could be viewed as a mental effort produced to process the quantity of information involved in the task.

Unlike the engagement index which is directly extracted from the EEG data, the EEG workload index was based on pre-trained predictive model [103]. This model was trained using a set of EEG data collected from a training phase during which a group of seventeen participants performed a set of brain training exercises. This training phase involved three different types of cognitive exercises, namely: digit span, reverse digit span and mental computation. The objective of these training exercises was to induce different levels of mental workload while collecting the learner's EEG data. The manipulation of the induced workload level was done by varying the difficulty level of the exercises: by increasing the number of the digits in the sequence to be recalled for digit span and reverse digit span, and the number of digits to be added or subtracted

for the mental computation exercises - we refer to [103] for more details on the procedure. After performing each difficulty level, the participants were asked to report their workload level using the subjective scale of NASA Task load index (NASA_TLX) [166]. Once this training phase was completed, the collected EEG raw data were cut into 1-second segments and multiplied by a Hamming window. A FFT was applied to transform each EEG segment into a spectral frequency and generate a set of 40 bins of 1 Hz ranging from 4 to 43 Hz (EEG pre-treated vectors). The dimensionality of the data was then reduced using a Principal Component Analysis (PCA) to 25 components (the score vectors). Next, a Gaussian Process Regression (GPR) algorithm with an exponential squared kernel and a Gaussian noise [269] was run in order to train a mental workload predictive model (the EEG workload index) from the normalized score vectors. Normalization was done by subtracting the mean and dividing by the standard deviation of all vectors. In order to reduce the training time of the predictive model, a faster version of GPR the local Gaussian Process Regression algorithm was used [139]. The evaluation of this model showed a correlation with the participants' subjective scores NASA_TLX reaching 82% (mean correlation 72%). This same approach was used within an intelligent tutoring system called (MENTOR) fully controlled by this workload index to automatically select the most adapted learning activity for the learner. The experimental results showed positive impact of using such index on learners' performance and satisfaction.

The reader may question about the reliability of such kind of neural metrics. Actually, many contributions have tackled the issue of predicting human behavior from neural metrics, e.g., [244, 184, 144], by collecting EEG data to detect cognitive interest, emotional engagement and decision making of consumers towards communication messages or advertisements in order to optimize them.

6.3 Experimental setting

This section details the experimental session we set up to analyze the relation between the emotions and the argumentation process. More precisely, we detail the protocol we have defined to guide the experimental setting, and the resulting datasets we have manually annotated in order to combine the arguments proposed in the debates with the detected emotions. Finally, we specify the hypotheses we aim at verifying in this experiment.

Protocol

The general goal of the experimental session is to investigate the relation (if any) holding between the emotions detected in the participants during a debate session and the argumentation flow of the debate itself. The idea is to associate the arguments and the relations among them to the participants' mental engagement and workload detected via the EEG headset, and the facial emotions identified via the Face Emotion Recognition tool.

More precisely, starting from an issue to be discussed provided by the moderators, e.g., *We have to ban animal testing*, the aim of the experiment is to collect the arguments proposed by the participants on the topic, as well as the relations among them (i.e., support or attack), and to associate such arguments/relations to the mental engagement and workload states and to the facial emotions expressed by the participants. During a post-processing phase on the collected data, we synchronize the arguments put forward by the different participants at time t with the emotional indexes we retrieved. Finally, we build the resulting bipolar argumentation graph of each debate, such that the resulting argumentation graphs are labelled with the source who has proposed each argument, and the emotional state of each participant at the time of the introduction of the argument in the discussion.

The first point to clarify in this experimental setting is the terminology. In this experiment, an *argument* is each single piece of text that is proposed by the participants in the debate. Typically, arguments have the goal to promote the opinion of the debater in the debate. Thus, an *opinion* in our setting represents the overall opinion of the debater about the issue to be debated. The opinion is promoted in the debate through arguments, that will support (if the opinions converge) or attack (otherwise) the arguments proposed in the debate by the other participants.

The experiment involves two kinds of persons:

- *Participant*: she is expected to provide her own opinion about the issue of the debate proposed by the moderators, and to argue with the other participants in order to make them understand the goodness of her viewpoint (in case of initial disagreement).[3]
- *Moderator*: she is expected to propose the initial issue to be discussed to the participants. In case of lack of active exchanges among the participants, the moderator is in charge of proposing pro and con arguments (with respect to the main issue) to reactivate the discussion.

The experimental setting of each debate is conceived as follows: there are 4 participants involved in each discussion group, and 2 moderators. Each participant is placed far from the other participants, even if they are in the same room, while moderators are placed in another room. Moderators interact with the participants uniquely through the debate platform, and the same holds for the interactions among participants. The language used for debating is English.

In order to provide an easy-to-use debate platform to the participants, without requiring from them any background knowledge, we decide to rely on a simple IRC network[4] as debate platform. The debate is anonymous and participants are visible to each others with their nicknames, e.g., *participant1*, while the moderators are visualized as *moderator1* and *moderator2*. Each participant has been provided with 1 laptop device equipped with internet access and a camera used to detect facial emotions. Moreover, each participant has been equipped with an EEG headset to detect the engagement and workload indexes. Moderators were equipped with a laptop only.

The procedure we follow for each debate is:

- Participants are firstly equipped with the EEG headset and the good connection of the headset is verified;
- Participants are familiarized with the debate platform;
- The debate starts - Participants take part into two debates each, about two different topics for a maximum of about 20 minutes for each debate:
  - The moderator(s) provides the debaters with the topic to be discussed;
  - The moderator(s) asks each participant to provide a general statement about his/her opinion on the topic;
  - Participants expose their opinion to the others;
  - Participants are asked to comment on the opinions expressed by the other participants;

[3]Note that, in this experimental scenario, we do not evaluate the connection between the emotions and persuasive argumentation. This analysis is out of the scope of this chapter and it is left for future research.

[4]`http://webchat.freenode.net/`

– If needed (no active debate among the participants), the moderator(s) posts an argument and asks for comments from the participants;

The variables measured in this experimental setting are the following: *(i)* engagement and workload indexes (measurement tool: EEG headset), and *(ii)* facial emotions, i.e., Neutral, Happy, Sad, Angry, Surprised, Scared and Disgusted (measurement tool: FaceReader).

The post-processing phase of the experimental setting involves the following steps:

- manual annotation of the support and attack relations holding between the arguments proposed in each discussion, following the methodology described in Section Dataset;
- manual annotation of the opinion of the participants at the beginning and at the end of the debates they participated in, and synchronization with the debriefing questionnaire data;
- synchronization of the argumentation (i.e., the arguments/relations proposed at time instant t) with the emotional indexes retrieved at time t using the EEG headset and FaceReader.

Participants The experiment was distributed over 6 sessions of 4 participants each; the first session was discarded due to a technical problem while collecting data. We had a total of 20 participants (7 women, 13 men), whose age range was from 22 to 35 years. All of them were students in a North American university, and all of them had good computer skills. Since not all of them were native English speakers, the use of the Google translate service was allowed. They have all signed an ethical agreement before proceeding to the experiment.

Participants have been asked to complete a short questionnaire about their viewpoints on the discussed topics. Thus, after each debate session, a debriefing phase was addressed. The questionnaire contained the following questions[5]:

- What was your starting opinion about the discussed topic before entering into the debate?
- What is your final opinion about the discussed topic after the debate?
- If you changed your mind, why (i.e., which was the argument(s) that has made you change your mind)?

These questions allowed us to *know* what is the opinion of the participants about the specific topics they debated about, without the need to infer it from the arguments they propose in the debates. The answers participants provided to these questions have been then used to correlate their opinions with the emotions they felt during the debates.

Finally, participants have been asked to fill in a questionnaire for Big Five personality traits. More precisely, participants filled in a questionnaire of 50 items of the kind:

- I get stressed out easily;
- I don't like to draw attention on myself;
- I spend time reflecting on things;

[5]Such material is available at `http://bit.ly/DebriefingData`.

- ...

where the possible values range over a typical five-level Likert scale: *Totally Disagree*, *Disagree*, *Neutral*, *Agree*, *Totally Agree*. Such information allowed us to extract the *OCEAN* personality dimensions, i.e.:

O Openness, Originality, Open-mindedness

C Conscientiousness, Control, Constraint

E Extraversion, Energy, Enthusiasm

A Agreeableness, Altruism, Affection

N Neuroticism, Negative Affectivity, Nervousness

These dimensions have been analyzed with respect to their correlation with the detected emotions of participants during the debates. More details about this analysis are provided in the Results Section.

Dataset

In this section, we describe the dataset of textual arguments we have created from the debates among the participants. The dataset is composed of four main layers: *(i)* the basic annotation of the arguments proposed in each debate (i.e. the annotation in xml of the debate flow downloaded from the debate platform); *(ii)* the annotation of the relations of support and attack among the arguments; *(iii)* starting from the basic annotation of the arguments, the annotation of each argument with the emotions felt by each participant involved in the debate; and *(iv)* starting from the basic annotation, the opinion of each participant about the debated topic at the beginning, in the middle and at the end of debate is extracted and annotated with its polarity. In the reminder of this section, we describe the annotation process of the four layers and the resulting inter-annotator agreement to ensure the reliability of the produced linguistic resource.

The *basic* dataset is composed of 598 different arguments proposed by the participants in 12 different debates. The debated issues and the number of arguments for each debate are reported in Table 6.1. We selected the topics of the debates among the set of popular discussions addressed in online debate platforms like iDebate[6] and DebateGraph[7].

The annotation (in `xml`) of this dataset is as follows: we have assigned to each debate a unique numerical *id*, and for each argument proposed in the debate we assign an *id* and we annotate who was the participant putting this argument on the table, and in which time interval the argument has been proposed. An example of basic annotation is provided below:

```
<debate id="1" title="Ban_Animal_Testing">
<argument id="1" debate_id="1" participant="mod"
    time-from="19:26" time-to="19:27">Welcome to
    the first debate! The topic of the first debate
    is that animal testing should be banned.</argu-
    ment>

<argument id="3" debate_id="1" participant="2"
    time-from="20:06" time-to="20:06">If we don't
    use animals in these tests, what could we use?
    </argument>
</debate>
```

[6] http://idebate.org/

[7] www.debategraph.org/

The second level of our dataset consists in the annotation of arguments pairs with the relation holding between them, i.e., support or attack. To create the dataset, for each debate of our experiment we apply the following procedure, detailed in Section 5.3:

1. the main issue (i.e., the issue of the debate proposed by the moderator) is considered as the starting argument;
2. each opinion is extracted and considered as an argument;
3. since *attack* and *support* are binary relations, the arguments are coupled with:
   - the starting argument, or
   - other arguments in the same discussion to which the most recent argument refers (e.g., when an argument proposed by a certain user supports or attacks an argument previously expressed by another user);
4. the resulting pairs of arguments are then tagged with the appropriate relation, i.e., *attack* or *support*.

To show a step-by-step application of the procedure, let us consider the debated issue *Ban Animal Testing*. At step 1, we consider the issue of the debate proposed by the moderator as the starting argument (a):

(a) *The topic of the first debate is that animal testing should be banned.*

Then, at step 2, we extract all the users opinions concerning this issue (both pro and con), e.g., (b), (c) and (d):

(b) *I don't think the animal testing should be banned, but researchers should reduce the pain to the animal.*

(c) *I totally agree with that.*

(d) *I think that using animals for different kind of experience is the only way to test the accuracy of the method or drugs. I cannot see any difference between using animals for this kind of purpose and eating their meat.*

(e) *Animals are not able to express the result of the medical treatment but humans can.*

At step 3a we couple the arguments (b) and (d) with the starting issue since they are directly linked with it, and at step 3b we couple argument (c) with argument (b), and argument (e) with argument (d) since they follow one another in the discussion. At step 4, the resulting pairs of arguments are then tagged by one annotator with the appropriate relation, i.e.: **(b)** *attacks* **(a)**, **(d)** *attacks* **(a)**, **(c)** *supports* **(b)** and **(e)** *attacks* **(d)**. For the purpose of validating our hypotheses, we decided to not annotate the supports/attacks between arguments proposed by the same participant (e.g., situations where participants are contradicting themselves). Note that this does not mean that we assume that such situations do not arise: no restriction was imposed to the participants of the debates, so situations where a participant attacked/supported her own arguments are represented in our dataset. We just decided to not annotate such cases in the dataset of argument pairs, as it was not necessary for verifying our assumptions.

To assess the validity of the annotation task and the reliability of the obtained dataset, the same annotation task has been independently carried out also by a second annotator, so as to compute inter-annotator agreement. It has been calculated on a sample of 100 argument pairs (randomly extracted). The complete percentage agreement on the full sample amounts to 91%. Applying such formula to our data, the inter-annotator agreement results in $\kappa = 0.82$. As a rule of thumb, this is a satisfactory agreement, therefore we consider these annotated datasets as reliable (i.e., our *goldstandard* dataset where arguments are associated to participants' emotions detected by EEG/FaceReader) to be exploited during the experimental phase.

Dataset				
Topic	**#arg**	**#pair**	**#att**	**#sup**
BAN ANIMAL TESTING	49	28	18	10
GO NUCLEAR	40	24	15	9
HOUSEWIVES SHOULD BE PAID	42	18	11	7
RELIGION DOES MORE HARM THAN GOOD	46	23	11	12
ADVERTISING IS HARMFUL	71	16	6	10
BULLIES ARE LEGALLY RESPONSIBLE	71	12	3	9
DISTRIBUTE CONDOMS IN SCHOOLS	68	27	11	16
ENCOURAGE FEWER PEOPLE TO GO TO THE UNIVERSITY	55	14	7	7
FEAR GOVERNMENT POWER OVER INTERNET	41	32	18	14
BAN PARTIAL BIRTH ABORTIONS	41	26	15	11
USE RACIAL PROFILING FOR AIRPORT SECURITY	31	10	1	9
CANNABIS SHOULD BE LEGALIZED	43	33	20	13
TOTAL	**598**	**263**	**136**	**127**

Table 6.1: The dataset of argument pairs resulting from the experiment.

Table 6.1 reports on the number of arguments and pairs we extracted applying the methodology described before to all the mentioned topics. In total, our dataset contains 598 different arguments and 263 argument pairs (127 expressing the *support* relation among the involved arguments, and 136 expressing the *attack* relation among the involved arguments).

The dataset resulting from these three layers of annotation adds to all previously annotated information the player characteristics (gender, age and personality type), FaceReader data (dominant emotion, Valence (pleasant/ unpleased) and Arousal (activated/ inactivated)), and EEG data (Mental Engagement levels)[8]. A correlation matrix has been generated to identify the correlations between arguments and emotions in the debates, and a data analysis is performed to determine the proportions of emotions for all participants. We consider the obtained dataset as the reference dataset to carry out our empirical study.

An example, from the debate about the topic "Religion does more harm than good" where arguments are annotated with emotions (i.e., the third layer of the annotation of the textual arguments we retrieved), is as follows:

[8]The datasets of textual arguments are available at `http://project.inria.fr/seempad/datasets/`.

```
<argument id="30" debate_id="4" participant="4"
time-from="20:43" time-to="20:43"
emotion_p1="neutral" emotion_p2="neutral"
emotion_p3="neutral" emotion_p4="neutral">
Indeed but there exist some advocates of the devil
like Bernard Levi who is decomposing arabic
countries. </argument>

<argument id="31" debate_id="4" participant="1"
time-from="20:43" time-to="20:43"
emotion_p1="angry" emotion_p2="neutral"
emotion_p3="angry" emotion_p4="disgusted">
I don't totally agree with you Participant2:
science and religion don't explain each other,
they tend to explain the world but in two
different ways.</argument>

<argument id="32" debate_id="4" participant="3"
time-from="20:44" time-to="20:44"
emotion_p1="angry" emotion_p2="happy"
emotion_p3="surprised" emotion_p4="angry">
Participant4: for recent wars ok but what
about wars happened 3 or 4 centuries ago?
</argument>
```

Finally, the fourth annotation task starts from the basic one, and it selects for each participant one argument at the beginning of the debate, one argument in the middle of the discussion, and one argument at the end of the debate. These arguments are then annotated with their *polarity* with respect to the issue of the debate: *negative*, *positive*, or *undecided*. The negative polarity is assigned to an argument when the opinion expressed in such argument is against the debated topic, while the positive polarity label is assigned when the argument expresses a viewpoint that is in favor of the debated issue. The undecided polarity is assigned when the argument does not express a precise opinion in favor or against the debated topic. Selected arguments are evaluated as the most representative arguments proposed by each participants to convey her own opinion, in the three distinct moments of the debate. The rationale behind this annotation is that it allows to easily detect when a participant has changed her mind with respect to the debated topic. An example is provided below from the debate "Ban partial birth abortions", where Participant4 starts the debate being undecided and then turns to be positive about banning partial birth abortions in the middle and at the end of the debate:

```
<argument id="5" participant="4" time-from="20:36"
time-to="20:36" polarity="undecided">Description's
gruesome but does the fetus fully lives at that
point and therefore, conscious of something ? Hard
to answer. If yes, I might have an hesitation to
accept it. If not, the woman is probably mature
enough to judge.</argument>

<argument id="24" participant="4" time-from="20:46"
time-to="20:46" polarity="positive">In the animal
world, malformed or sick babies are systematically
abandoned.</argument>

<argument id="38" participant="4" time-from="20:52"
time-to="20:52" polarity="positive">Abortion is
legal and it doesn't matter much when and how.
It's an individual choice for whatever reason
it might be.</argument>
```

Hypotheses

The experiment we have carried out aims at verifying the link between the emotions detected on the participants of the debate, and the arguments and their relations proposed in the debate. Our hypotheses therefore revolve around the assumption that the participants' emotions arise out of the arguments they propose in the debate:

H1 : The argumentation process in a debate requires high mental engagement and generates negative emotions when the interlocutor's arguments are attacked.

H2 : The number of arguments and attacks proposed by the debaters are correlated with negative emotions.

H3 : The number of expressed arguments is connected to the degree of mental engagement and social interactions.

H4 : The personality of the participants modulates their emotional experiences during the debates.

H5 : The debaters' opinions regarding the discussed topics have an impact on their emotions.

6.4 Results

In order to verify the above mentioned hypotheses, we first computed the mean percentage of appearance of each basic emotion across the 20 participants. Results show (with 95% of confidence interval) that the most frequent emotion expressed by participants was *anger*, with a mean appearance frequency ranging from 8.15% to 15.6% of the times. The second most frequent emotion was another negative emotion, namely *disgust*, which was present 7.52% to 14.8% of the times. The overall appearance frequency of other emotions was very low. For example, the frequency of appearance of happiness was below 1%. Even if this result might be surprising at a first glance, this trend can be justified by a phenomenon called *negativity effect* [279]. This means that negative emotions have generally more impact on a person's behavior and cognition than positive ones. So, negative emotions like anger and disgust have a tendency to last in time more than positive emotions like happiness.

With regard to the mental engagement, participants show in general a high level of attention and vigilance in 70.2% to 87.7% of the times. This high level of engagement is also correlated with appearance of anger (r=0.306), where r refers to the Pearson product-moment correlation coefficient. This coefficient is a standard measure of the linear correlation between two variables X and Y, giving a value between $[1,-1]$, where 1 is a total positive correlation, 0 means no correlation, and -1 is a total negative correlation. This trend confirms that, in such context, participants may be thwarted by the other participant arguments or attacks, thus the level of engagement tends to be high as more attention is allowed to evaluate the other arguments or to formulate rebutting or offensive arguments. Thus, our experiments confirm behavioral trends as expected by the first hypothesis.

Figure 6.2 shows an evolution of the first participant's emotions at the beginning of the first debate. The most significant lines of emotions are surprise and disgust (respectively, the line with squares and the line with circles). The participant is initially surprised by the discussion (and so mentally engaged) and then, after the debate starts, this surprise switches suddenly into disgust, due to the impact of the rejection of one of her arguments; the bottom line with circles grows and replaces the surprise as the participant is now actively engaged in an opposed argument (thus confirming our hypothesis 2). Finally, the participant is calming down. In this line, Figure 6.3 highlights that we have a strong correlation (r= 0.83) in Session 2

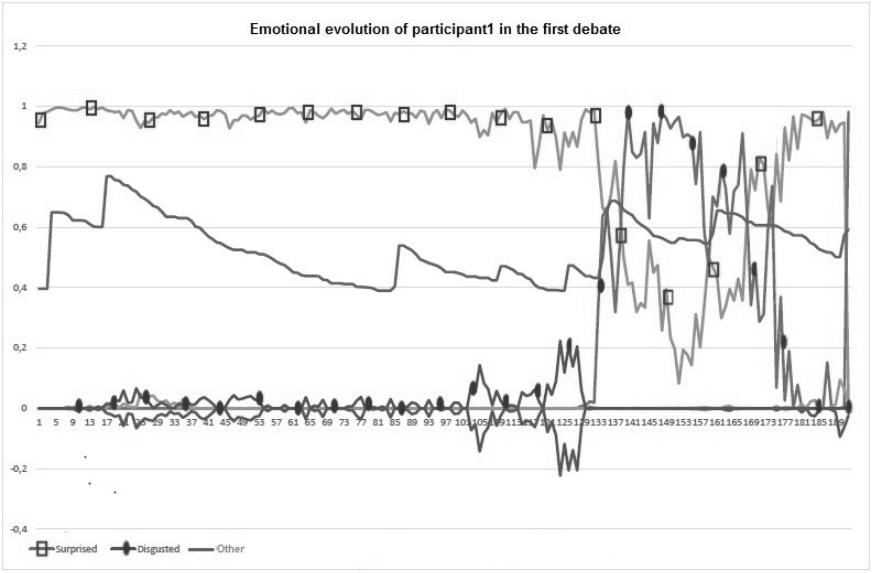


Figure 6.2: Emotional evolution of Participant 1 in Debate 1 (lines with squares and circles represent, respectively, the *surprise* and *disgust* emotions.

	NB ARG	ATTACK	SUPPORT
Pleasant	0,0962	0,1328	-0,0332
Unpleasant	-0,0962	-0,1328	0,0332
High ENG	-0,0718	**-0,6705**	0,2459
LowENG	-0,2448	0,2115	-0,1063
Neutral	0,0378	0,6173	-0,1138
Disgusted	-0,0580	**-0,4367**	-0,3621
Scared	0,1396	-0,0952	**0,5755**
Angry	-0,1018	-0,4386	0,0582

Figure 6.3: Correlation table for Session 2 (debated topics: *Advertising is harmful* and *Bullies are legally responsible*.

showing that the number of attacks provided in the debate increased linearly with the manifestation of more disgust emotion.

In the second part of our study, we were interested in analyzing how emotions correlate with the number of attacks, supports and arguments. We have generated a correlation matrix to identify the existent correlations between arguments and emotions in debates. Main results show that the number of arguments tends to decrease linearly with manifestations of sadness (r=-0.25). So when the participants start to feel unpleasant emotions, such as sadness, the number of arguments decreases, showing a less positive social behavior[9] and a tendency to retreat into herself. This negative correlation between the number of arguments and sadness

[9]By positive social behavior, we mean that a participant aims at sharing her arguments with the other participants. This attitude is mitigated if unpleasant emotions start to be felt by the participant.

	NB ARG	ATTACK	SUPPORT
Pleasant	**0,7067**	-0,3383	-0,3800
Unpleasant	-0,7067	0,3383	0,3800
High ENG	**-0,6903**	-0,3699	-0,1117
LowENG	-0,1705	**0,5337**	-0,0615
Neutral	0,8887	-0,0895	-0,3739
Disgusted	0,1017	**0,8379**	**0,5227**
Scared	0,2606	-0,4132	**-0,7107**
Angry	**-0,7384**	-0,5072	-0,0937

Figure 6.4: Correlation table for Session 3 (debated topics: *Distribute condoms at schools* and *Encourage fewer people to go to the university*).

	NB ARG	ATTACK	SUPPORT
Pleasant	**0,1534**	0,0134	-0,0493
Unpleasant	-0,1534	-0,0134	0,0493
High ENG	-0,0246	-0,0437	**0,3185**
LowENG	**0,2054**	0,1147	0,1592
Neutral	0,0505	0,1221	-0,2542
Disgusted	-0,0177	-0,0240	**0,2996**
Scared	-0,0278	0,0297	-0,2358
Angry	0,0344	**-0,2206**	0,0782

Figure 6.5: General correlation table of the results.

even reaches very high level in certain debates (i.e. a mean correlation r= -0.70 is registered in the two debates of the second session). Another negative linear relationship is registered with regard to the number of attacks and the anger expressed by the participant (r=0.22). Participants who tend to attack the others in the debate are less angry than those whose number of attacks is smaller. Figure 6.4 shows the correlation table for Session 3. The analysis of the results we obtained shows the occurrence of strong correlations between emotions and attacks / media / number of arguments in some discussions, but not in others. This is an interesting index to investigate in future work.

Figure 6.5 shows the most significant correlations we detected. For instance, the number of supports provided in the debate increased linearly with the manifestation of high levels of mental engagement (r=0.31). This trend is more pronounced when the debate does not trigger controversies and conflicts between the participants. For example, in the debate *Encourage fewer people to go to university*, all the participants shared the same opinion (against the main issue as formulated by the moderator) and engaged to support each other's arguments. The correlation between the number of supports and the engagement was very high (r=0.80) in this debate. The number of attacks is more related to low engagement. The moderator can provide more supporting arguments to balance participants' engagement, and if the attacks are increasing, that means participants tend to disengage. The experiments show that participants maintaining high levels of vigilance are the most participative in the debate and resulted in a more positive social behavior (thus confirming our hypothesis 3).

Our next objective was to check whether participants' emotions during the debates were modulated by their personality. In other words, we wanted to see whether there was any impact of the debaters' personality

on their emotional responses in terms of facial expressions, valence, engagement and cognitive load indexes. The Big Five inventory data [180] were considered for this analysis. Participants were classified according to each of the five OCEAN personality dimensions, namely *openness* (imaginative vs. more pragmatic participants), *conscientiousness* (conscientious vs. non conscientious), *extroversion* (extroverted vs. introverted), *agreeableness* (compassionate toward others vs. more antagonistic), and *neuroticism* (anxious vs. emotionally stable). Multivariate analyses of variance (MANOVA) were run with the OCEAN personality traits as fixed factors and the debaters' emotions as dependent variables. These included the following combined measures: (1) the proportions of occurrences of the six facial expressions, i.e. happiness, anger, fear, sadness, disgust, and surprise, (2) the proportions of negative and positive emotional valence, (3) the proportions of high, medium and low levels of mental engagement, and (4) the proportions of high, medium and low workload. In total, 20 MANOVA were conducted, crossing the 5 personality traits with the 4 dependent variables.

Three statistically reliable MANOVA were found, showing significant relationships between the debaters' personality traits and emotional responses[10]:

- Extroversion and facial expressions ($F(6, 33) = 2.574$, $p < 0.05$, Pillai's Trace = 0.319). In particular, extroverted participants showed significantly more frequently expressions of surprise than the introverted participants (M = 6.70%, SE = 1.10% vs. M = 1.70%, SE = 1.30%; $F(1, 38) = 8.385$, $p < 0.008$). This can be explained by the fact that introverted people tend to hide their emotions as compared to extroverted people.

- Conscientiousness and workload ($F(3, 36) = 5.200$, $p < 0.05$, Pillai's Trace = 0.302). More precisely, participants with a non conscientious temperament had significantly more occurrences of low levels of workload as compared to the other participants (M = 29.6%, SE = 3.8% vs. M = 16.0%, SE = 3.8%; $F(1, 38) = 6.525$, $p < 0.016$). They seemed to experience on average less cognitive load during the discussions.

- Neuroticism and mental engagement ($F(3, 36) = 3.518$, $p < 0.05$, Pillai's Trace = 0.227). In particular, participants with an anxious temperament had on average significantly fewer proportions of high engagement levels during the debates as compared to the other participants (M = 18.0%, SE = 2.4% vs. M = 28.5%, SE = 3.0%; $F(1, 38) = 7.423$, $p < 0.016$) . This can be seen as follows: anxious people tend to be easily stressed. They are thus likely to have trouble concentrating, and hence have difficulties being mentally engaged, as opposed to less emotionally vulnerable people.

To summarize, these results validate our fourth hypothesis: the personality has an important impact on the debaters' emotional responses. Inner emotions (brain activity) seem to be modulated by the neuroticism and the conscientiousness temperament traits. Outer emotions (facial expressions) were modulated by the extroversion traits. Neuroticism and conscientiousness have both a negative impact on the debaters' brain indexes, with respectively, a reduced mental engagement index and an increased cognitive load. For facial expressions, we have particularly found that the emotion of surprise was more frequent among the debaters with an extroverted temperament. This is an important aspect considering that this expression was the least observed during our experiments. Indeed when analyzing the debaters' emotions with FaceReader, we observed that the expression of surprise was hardly dominant (compared with neutral) during the discussions.

[10] A Bonferroni correction (.05 divided by the number of dependent variables) has been applied within the MANOVA follow-up analyses to account for multiple ANOVAs being run.

The dominance of the other facial expressions, namely anger, fear, sadness and disgust does not seem to be influenced by the participants' personality.

Our next concern was to investigate if there were any differences in terms of emotional experiences (facial expressions, emotional valence, mental engagement and workload) during the debates according to the participants' opinions on the discussed topics. These opinions were given during the debriefing as previously mentioned. Each participant was either for or against the topic of the debate: *for* means that the participant agreed with the subject of the debate (e.g. for distributing condoms to students), and *against* means that the participant disagreed with the addressed topic (e.g. against distributing condoms in schools). A participant could also have no particular opinion (*no-opinion*) regarding the topic of the debate if he was neither for nor against. Moreover, each participant was asked to give a starting opinion, before the discussion, and a final opinion, after the debate. The goal was to assess the impact, if any, of changes in opinions on the debaters' emotions. Table 6.2 enumerates participants' initial and final opinions.

Starting/Final	No-opinion	For	Against	*Total*
No-opinion	**2**	5	0	7
For	0	**12**	1	13
Against	0	1	**19**	20
Total	2	18	20	40

Table 6.2: Number of opinions before and after the debates (in bold the number of debaters who kept the same opinion).

As for the previous hypothesis, distinct MANOVA were performed to analyze the proportions of occurrence of *(1)* facial expressions: happy, sad, angry, surprised, scared and disgusted; *(2)* valences of emotions: positive and negative; *(3)* engagement levels: high, medium and low; and *(4)* cognitive load levels: high, medium and low. First, we wanted to check whether there were any significant differences in terms of emotions between the participants who kept the same opinion during the debates (N = 33) and the participants who changed their opinion (N = 7) . Then, for those who kept the same opinion, we wanted to compare the emotional responses between the participants who were for and the participants who were against[11].

No statistically reliable effect was found in any of the performed analyses (p = n.s.) suggesting that there were no significant differences in terms of facial expressions, valence, engagement and workload, neither between the debaters who kept the same opinions and the debaters who changed their opinion, nor between those who were for and those who were against the discussed topics throughout the debates.

In addition to these analyses, the debaters' starting and final opinions were studied independently. That is, we checked whether either the former or the latter opinions had (independently of each other) an impact on the emotions expressed during the debates. Again no statistically significant effect was found. This has led us to conclude that neither the initial nor the final opinions had an impact on the debaters' emotional states. To summarize, the emotional experience during the debates does not seem to be related to the opinion of the debaters regarding the addressed topics. Emotions are rather depending on the person's temperament and the dynamics of the debate (i.e. arguments, supports and attacks).

[11]The two participants who did not have an opinion throughout the debate were discarded since they have reported they could not follow the discussions because of their lack of understanding of English.

Examples of correlations on single debates

In this section we provide some examples of correlations among the emotions and the argumentative elements emerging from the single debates. The goal is to provide a more detailed analysis, given the fact that some debates have been more passionate than others because of the personal involvement of the participants in the subject of the debate. It is worth noticing that as the number of instances involved in our debates is 4, these correlations cannot be considered as significative. However, we believe that these examples may show interesting insights to be investigated in our future experiments. The categories of correlations we investigate are the following: engagement vs argumentation; engagement vs emotions; workload vs argumentation; workload vs emotions; pleasant/unpleasant vs argumentation; Big5 vs emotions. Correlation values are comprised between -1 (negative correlation) and +1 (positive correlation). In our analysis we consider as strong correlations the values between -0.7 and -1 (strong negative correlation), and between 0.7 and 1 (strong positive correlation). All values in between cannot be considered significant to verify our hypothesis.

Debate: Advertising is harmful Number of arguments in the debate: 71 (64 from participants and 7 from moderators).

Workload vs emotions: The more the participants feel surprised, the higher the workload is high ($r = 0.80$) meaning that the mental load increases if the participants feel surprised about the arguments proposed in the ongoing debate. Moreover, the more the participants feel neutral, the lower the workload ($r = 0.82$) meaning that the mental load decreases if the participants feel neutral with respect to the ongoing debate, i.e., they are not interested in the topic of the debate as well as in the the other participants' arguments.

Big5 vs emotions: Participants with a high degree of *agreeableness* are more inclined to be surprised ($r = 0.90$), while participants with a high degree of *conscientiousness* tend to get sad or angry if the debate is not going in the desired direction (correlations $r = 0.82$ and $r = 78$, respectively), since they are inclined to do their duty well and thoroughly.

Debate: Students are legally responsible for bulling Number of arguments in the debate: 71 (66 from participants and 5 from moderators).

Engagement vs emotions: On the one side, we have a strong correlation between the high engagement and the emotion *happy* ($r = 0.94$), meaning that when the participants of this debate are experiencing such positive emotion they become more passionate (and therefore engaged) in the discussed topic. On the other side, we have also a strong correlation between the low engagement and the emotion *disgusted* ($r = 0.82$), meaning that when participants experience such negative emotion, then they become less interested in the ongoing debate.

Pleasant/Unpleasant vs argumentation: Strong correlation between positive valence (pleasant) and the number of arguments proposed in the debate ($r = 0.74$), meaning that when participants propose more arguments in the debate, they are more interested in the debated topic and therefore there is a higher degree of pleasantness in the air.

Big5 vs emotions: Participants with a high degree of *extroversion* or of *agreeableness* are more inclined to externalize that they fell surprised about the ongoing debate (correlations $r = 0.96$ and $r = 89$, respectively). Moreover, participants with a high degree of *neuroticism* are more inclined to be disgusted about the arguments proposed in the debate ($r = 0.82$).

Debate: Distribute condoms in schools Number of arguments in the debate: 68 (63 from participants and 5 from moderators).

Engagement vs emotions: Strong correlation between the high engagement and the emotion *angry* ($r = 0.96$), meaning that when the participants are experiencing such negative emotion they become more passionate (and therefore engaged) in the discussed topic.

Workload vs argumentation: Strong correlation between a high degree of workload and the number of supports among the arguments ($r = 0.86$), meaning that when the number of supports increases then the participants of this debate are required with a higher mental load to understand how the debate is going on.

Workload vs emotions: The more the participants feel disgusted, the higher the workload is ($r = 0.92$) meaning that the mental load increases if the participants feel disgusted about the arguments proposed in the ongoing debate, as they need to construct in their minds new arguments to defeat the ones proposed by the other participants that make them feel disgusted.

Big5 vs emotions: Participants with a high degree of *extroversion* are more inclined to externalize that they fell surprised about the ongoing debate ($r = 0.77$). Moreover, participants with a high degree of *neuroticism* are more inclined to be angry about the arguments proposed in the debate ($r = 0.75$).

Debate: We should fear the power of government over the internet Number of arguments in the debate: 41 (37 from participants and 4 from moderators).

Engagement vs emotions: Strong correlation between the high engagement and the emotion *disgusted* ($r = 0.93$), meaning that when the participants are experiencing such negative emotion they become more passionate (and therefore engaged) in the discussed topic.

Workload vs argumentation: Strong correlation between a low degree of workload and the number of supports among the arguments ($r = 0.86$), meaning that when the number of supports increases participants require a lower mental load to understand how the debate is going on.

Workload vs emotions: The more the participants feel *angry*, the lower the workload is ($r = 0.93$). This means that those participants that become angry due to the arguments that are proposed in the ongoing debate tend to use less mental resources to propose new, possibly effective, arguments.

Pleasant/Unpleasant vs argumentation: Strong correlation between negative valence (unpleasant) and the number of attacks between arguments ($r = 0.70$), meaning that when participants disagree attacking each others there is an higher degree of unpleasantness in the air.

Big5 vs emotions: Participants with a high degree of *extroversion* are more inclined to externalize that they feel happy about the ongoing debate ($r = 0.99$). Moreover, participants with a high degree of *neuroticism* are more inclined to be disgusted about the arguments proposed in the debate ($r = 0.90$).

Note that this study is dealing with correlation between the nature of the arguments and their relations (support/attack) and the participants' emotions, and we are not claiming to have found a direct causal relation between the arguments and such users' emotions - which is out of the scope of this current study. It is however an interesting direction for further work with larger sample size using [157]'s causality test.

Discussion

We have learnt several lessons from the realized experiment. First, the different debates and participants have confirmed the correctness of our hypotheses. Debates constitute the underlying framework for generating emotions which evolve with the argumentation flow. The difference of opinions is the starting point of the rise and successive transformation of specific emotions. However, so far we have not taken into account the initial emotional state of the participants (i.e. before starting the discussion on the topic), that can influence the participants reactions during the debate. We will have to consider this in further studies.

Facial emotion recognition and EEG measures allowed us to identify not only the type of emotion generated, but also the intensity and the evolution of emotions. Associated with the workload index, this also allowed us to detect how the participant is engaged in the discussion and so, how he holds on to his arguments. Being strongly convinced by an opinion provokes the birth of a mental energy strong enough to increase the workload and develop a justification. Contradictions with the flow of arguments generate anger which evolves progressively into disgust if the participant's arguments apparently cannot convince the opponents. In the classification of emotion, disgust (which is close in terms of emotion) is a normal evolution of anger and appears when the participant feels a dual feeling for two reasons: 1) he is not pleased with himself for not having convinced the opponent (internal feeling), and 2) he has a very low opinion of the opponent (external feeling). We highlighted the evolution of this emotion in several debates showing the important consequence of the argumentation by provoking internal evolution of emotions. This can be explained by the impact of a contradiction on an in depth conviction. The more a participant is convinced of the merits of his position, the more he will be subject to a strong emotion.

The three dimensions of our evaluation framework (emotion recognition, engagement and workload) allowed us to assess more precisely the impact of argumentation on emotional response throughout the debate. Workload decreases when participants feel angry, which means that they reduce their ability to use or construct new arguments. When this emotion evolves to disgust, the workload increases which means that they have to reconsider their own set of arguments either for an update or a new construction. High engagement provokes the rise of strong positive or negative emotions while, on the other side, we have confirmed that disgusted participants become less engaged in the ongoing debate. Finally, we have considered the influence of personality to the type of generated emotion. Participants with a high degree of neuroticism are converging to be angry or disgusted, which are close emotions. Participants with a high degree of agreeableness or extroversion are more open to feel surprise.

Note that the goal of this experiment is not to learn how to best intervene to improve online discourse but to study what are the insights that online cognitive agents and bots need to implement to address dialogues with humans. A cognitive agent, in order to behave like humans in debates, has to feel emotions and generate them in the other agents (being them humans or artificial) that interact with it. This extensive study is the first but essential step towards a better comprehension of the relation between human emotions and argumentation. As a shorter term objective, the aim of this contribution is to guide the definition of the next argumentation frameworks such that not only objective elements are taken into account but also cognitive ones.

From this experiment, we learnt that argumentation in online debates cannot be considered as a standalone process, as it discloses many emotional aspects, e.g., when users are more engaged in a discussion more arguments are proposed, and the most engaging discussions are correlated with negative emotions like anger and disgust. Moreover, a strong correlation exists among personality traits and the emotions felt by participants during online argumentation, e.g., the dominance of emotions like anger, fear, sadness and disgust does not seem to be influenced by the participants' personality where emotions of happiness and surprise were more frequent among the debaters with an extroverted temperament.

6.5 Argumentative persuasion and emotions in humans: a field experiment

In everyday life situations like online discussions and political debates, "the aim of persuasion is for the persuader to change the mind of the persuadee" [173]. This process, called *persuasive argumentation*, may employ different strategies. In the Ethos strategy, persuasion relies on the authority of the persuader with respect to the topic of the debate. The Logos strategy is grounded on logical arguments leading to a sound inference process to derive conclusions, while the Pathos strategy solicits the emotions of the interlocutors to generate empathy. These strategies have been used to define formal models of persuasion, e.g., [173], to be employed by intelligent agents to persuade the others to change their beliefs. However, analyzing how these strategies are perceived by humans when they argue, and what is the impact of these strategies on the humans' mental states like *engagement* and *emotions* has not been explored. Yet, this would be of valuable importance for argumentative agents to be able to apply persuasion strategies as humans do, resulting in more effective interactions with people.

In this chapter, we answer the following research question: *what is the impact of persuasion strategies on the mental states and emotions of the debaters?*

Three kinds of argumentative persuasion exist: *Ethos*, *Logos*, and *Pathos* [277]. Ethos deals with the character of the speaker, whose intent is to appear credible. The main influencing factors for Ethos encompass elements such as vocabulary, and social aspects like rank or popularity. Logos is the appeal to logical reason: the speaker wants to present an argument that appears to be sound to the audience. Pathos encompasses the emotional influence on the audience.

In this chapter, we investigate also the distribution of engagement among the brain lobes [297, 313]: the *Frontal lobe* has two key functions, i.e., controlling motor activities (including speech), and human "executive functions" (e.g., planning, reasoning, making decision); the *Temporal lobe* controls visual and auditory memories; the *Parietal lobe* is responsible for processing sensory information, comprehending oral and writing language, and controlling working memory; the *Occipital lobe* is responsible for vision.

We conducted a field experiment with users, starting from three hypotheses to be validated. We raised a number of debates in which, together with the participants of the experiment, a *persuader* was involved to convince the other participants about the goodness of her viewpoint, applying one of the three persuasion strategies. The persuader is a person who has been provided with particular argumentation frameworks but appears to the other participants as just another participant, e.g., she does not dominate the debate. Every participant was equipped with an Electroencephalography (EEG) Headset to detect mental engagement, and cameras to detect facial emotions. The collected data was synchronized to assess the validity of our hypotheses. Results highlight the higher persuasion impact of the Pathos strategy.

We combined physiological sensors (EEG) with facial expression analysis system (FaceReader 6.1).[12] By analyzing the user's face streamed via webcam, the FaceReader software is able to recognize six basic

[12] `www.noldus.com/human-behavior-research/products/facereader`

emotions: happy, sad, angry, surprised, scared and disgusted. The FaceReader model reaches 87% accuracy by extracting and classifying in real-time 500 key points in facial muscles. As output, FaceReader provides the probability of the presence of these six emotions, as well as the probability of the neutral state.

Experimental setting

The goal of our experiment was to investigate how the argumentative persuasion process in debates is affected by the mental states and emotions of the participants, and vice-versa. In each debate, besides the participants equipped with the EEG Emotiv EPOC devices, there is a participant who plays the role of the *persuader*, called the PP in the remainder of the chapter. The PP adopts and maintains a predefined viewpoint in the debate (i.e., pro or con), together with an argumentation strategy (i.e., Logos, Pathos or Ethos). PP intends to persuade other debaters of her viewpoint on the debated issue. The goal is to evaluate the following hypotheses:

H1: Argumentation strategies trigger negative emotions and engagement having an impact on the persuasion.

H2: Specific brain lobes are activated when a Logos or an Ethos argument is proposed by the PP, while other lobes are solicited when the PP puts forward a Pathos argument.

H3: Pathos arguments activate a higher empathy, triggering a number of arguments put forward by the other participants to support PP's arguments. Pathos arguments have a more effective persuasive power in the debate.

Participants and roles. 4 participants aged from 19 to 45 were involved in each of the 5 debate sessions, and each participant received a compensation of 20$ at the end of the session. In total, we collected data from 20 participants (7 women, 13 men). The size of the experiment is driven by the complexity of the experimental setting (devices, protocol). Debaters were preselected after filling an online form that collects their initial opinions about all the debate subjects, data is anonymized and kept confidential. This step was necessary to ensure possibly conflicting initial opinions in the debates. The ideal configuration includes 2 participants in favor and 2 against the debated topic. When not possible, a random assignment has been carried out. Each participant was kept separate from the others to avoid interactions out of the debate platform. In addition to the four participants and the PP, a *moderator* who proposes the debated issue and solicits unresponsive participants participated too. Each group of participants was involved in two debates. All participants (including the PP) were identified in the debate platform through a nickname. The PP cannot be identified by her nickname. No personal information about participants was disclosed during the debates.

Protocol. *Phase 0*: Participants fill in the self-reporting questionnaire about their initial opinions on the debate topics. They are associated to the debate sessions.
Phase 1: Familiarization of the participants with the Internet Relay Chat debate platform, the EEG headset, the camera for emotion recognition, and signature of a consent form.
Phase 2: The debate starts. Participants are involved in two debates for a maximum of 20 minutes each. The moderator provides the debaters with the topic to be discussed, and asks each participant to provide a general statement about her opinion on the topic. Each participant writes her viewpoint to the others, then the others are asked to comment on the expressed opinions. The PP plays the predefined persuasion strategy to convince the others with a different opinion, meaning that all arguments put forward by the persuader apply only the selected strategy. No turn taking was applied. Participants were free to propose their arguments,

and the PP participates in the debate with the same amount of arguments as the other participants. The debaters were free to put forward generic arguments about the debated topic, or to explicitly refer to the other participants' argument to attack or support them. Arguments proposed by the PP were pre-instantiated arguments retrieved on online debate platforms[13], and categorized with the three persuasion strategies we identified. These arguments allowed us to provide a fixed stimulus in the debate. When necessary, the PP slightly adapted the pre-defined argument to precisely refer to another participant's argument, e.g., "I don't agree with you Participant1 because `predefined argument`". After about 15 minutes of debate, the moderator asked to provide their final viewpoint on the topic, and the debate is closed. Strategies have not been randomized. For each debate session, the PP applies the logos strategy for one debate, and either Pathos or Ethos for the second debate to compare for each set of debaters a more rational strategy (i.e., Logos) vs a more empathic one (either Ethos or Pathos). The contingency table below shows the correlation of the strategy adopted by the persuader and her stance in the 10 debates.[14]
Phase 3: Participants are asked to fill a second self-reporting questionnaire on their experience in the debate.

Post-processing phase. we synchronized the textual argument collected during the debates, with the engagement index and the emotions.

	Stance		
Strategy	Pro	Con	Total by Strategy
Pathos	0	3	3
Logos	4	1	5
Ethos	1	1	2
Total by Stance	5	5	10

We are aware that field experiments, as the one proposed in this chapter, suffer from the possibility of contamination, and we agree about the fact that experimental conditions can be controlled with more precision in a constrained experimental setting. However, field experiments have the advantage that outcomes are observed in a natural setting rather than in a contrived environment, thus showing higher external validity than "laboratory" experiments. For instance, the reader may argue about our choice of an experimental setting where 5 persons are involved at the same time, instead of a more controlled setting with a 1:1 face-to-face exchange. However, our interest is not in studying the effect of a single strategy on a single person with respect to a single dialogue move, but in considering a more realistic setting where several persons interact, like on social media.

Dataset. Two annotation tasks have been carried out offline on the collected data[15] by two annotators. Each argument is annotated with debate identifier, argument identifier, participant, and timestamp.In total, 791 arguments, and 162 argument pairs (74 linked by an attack and 88 by a support) were annotated. We

[13] `www.debate.org/`, `www.createdebate.com/`

[14] The Pathos strategy has not been used with a Pro stance because *i)* we had 6 debate sessions but the EEG data of the first session, where we considered Pathos/Pro, was corrupted, and *ii)* the stance depends also on the arguments used on the debate platforms we collected to construct PP's ones.

[15] The corpus is available at `https://goo.gl/xSykTi`.

computed the inter-annotator agreement for the relation annotation task on 1/3 of the pairs of the dataset (54 randomly extracted pairs), obtaining a satisfactory agreement: $\kappa = 0.83$.

Experimental results

This section reports on the obtained results for our hypotheses. We divided the debate into three phases: the introduction (INTRO) where the PP states her own opinion on the topic of the debate; the argumentation (ARG) includes the reformulation, the refutation and the contribution of new ideas according to the strategy adopted by the PP; the conclusion (CONC) where the PP recalls her position and final opinion. This structure is inspired from the conversation structure in pragmatics, where conversations have a linear structure, i.e., initiation, maintenance and termination [187]. For data synchronization, we considered the participants' physiological reactions during 10 seconds after each intervention of the PP [195], and we computed the average emotion values of the 10 seconds after each argument proposal. We considered the anger scores in the result analysis because it was the most predominant emotion during the debates [278].

There is a significant correlation between the persuasion strategy and the participants' emotions	NO	H1
Engagement in supporters and anger in opponents grow in an inversely proportional way	YES	H1
Logos activates language comprehension and situations correlation	YES	H2
Logos activates planning and decision making	NO	H2
Ethos leads to the higher percentage of attacks wrt. PP's arguments	YES	H3
Pathos leads to the higher percentage of supports wrt. PP's arguments	YES	H3

Table 6.3: Experiments finding at a glance.

H1 - Persuasion vs. emotions and engagement. In this first hypothesis, we verified for each strategy, the means of anger generated throughout the different phases of the debate. To verify the impact of anger and engagement on persuasion, we ran a repeated ANOVA measure. As within-subjects factors, we consider the debate phases (INTRO, ARG, CONC). As between-subjects factors, we consider *PP_strategy* (Ethos, Logos, Pathos), measure (anger, engagement), and participant's final position (Neutral, Opponent, Supporter). We validate the repeated ANOVA measures with [218] test for sphericity on the dependent variable *Deb_phases* (sig=.013) (we assess the significance of the corresponding F with [159]'s correction). For the within-subject effect test, we have a significant effect of debate phases and *PP_strategy* on measuring (engagement and anger) with p=0.016 and F(8.857, 113.372)=2.405. The between-subject effects results show that there are significant main effects of the $PP\_strategy * Final\_Position$, F(8,64)=2.178, p=0.041, meaning a significant effect of the persuasion strategy, anger and engagement on persuasion. Fig. 6.6 presents the corresponding engagement to compare the effect of emotions on the engagement. Note that if anger decreases, the engagement increases in all persuasion strategies.

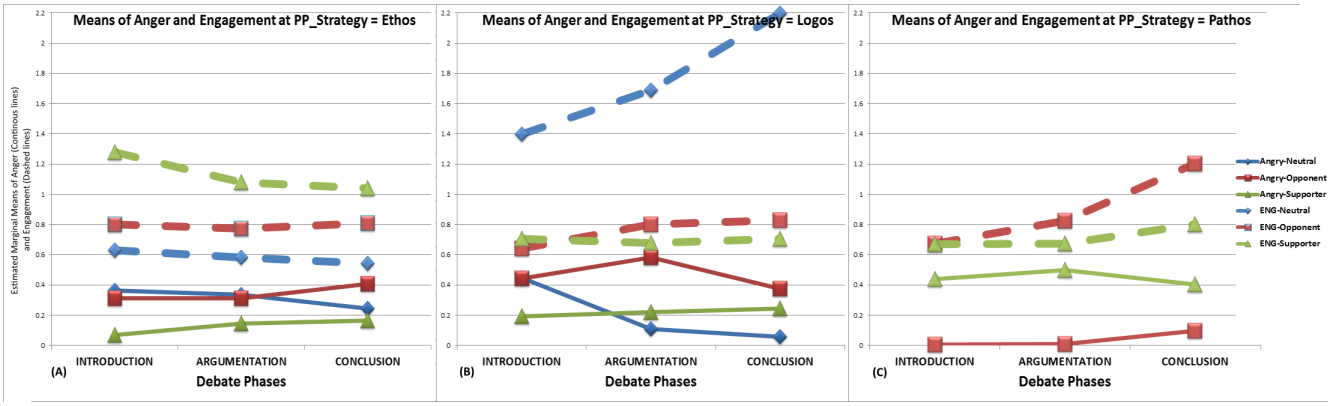


Figure 6.6: Means of anger (continuous lines) and engagement (dashed lines) (y axis) by debates' phases (x axis) for the different persuasion strategies. Blue, red and green colors correspond, respectively, to the participants' final position (Neutral, Opponent, and Supporter) to PP's opinion.

For the Logos strategy (Fig. 6.6-B), participants who stayed *Neutral* all over the debates had low negative emotions and their engagement was high. So participants who have not decided about the PP's opinion were more engaged in looking for logical reasons to support opinions. This can be interpreted as follows: neutral participants follow the arguments deployed by Logos and show a high engagement in trying to be persuaded. The opponents show a clear increase of negative emotions and loss of engagement. They are more engaged in the ARG phase in refuting the PP's arguments (*emotional resistance*) whereas the supporters were less engaged because they already accepted PP's logic. Hence, for the Logos strategy, neutral participants show decreasing negative emotions and engagement growth, whereas opponents are mostly subject to negative emotions and disengaged to follow the logical reasoning.

For the Ethos strategy (Fig. 6.6-A), opponents rejected the credibility of the PP and were not engaged in following her opinion. Their position does not change during the debates end where the negative emotion is higher. The neutrals were less engaged throughout the debate phases compared to the other participants. This can be due to the lack of interest in the subject of the debate and even disengagement in taking a position face to an expert opinion. We may notice that the supporters' engagement is higher in the INTRO phase, and continues to decrease at the ARG and CONC phases while their negative emotion is the lowest through the debate phases compared to other participants, indicating their satisfaction towards the expert's opinion.

For the Pathos strategy (Fig. 6.6-C), there are no neutral participants. We have opponents with increasing engagement related to the resistance to the emotional examples proposed by the PP. They were suppressing their negative emotion elicited by the Pathos strategy so their anger is low. Supporters were affected by Pathos, so their negative emotions are higher and their engagement is lower compared to the opponents because of the emotional effect of this strategy.

H2 - Brain solicitation vs. strategies. To verify the second hypothesis, we compute the differences in terms of engagement of each brain region for each participant, running a repeated ANOVA measure. The goal is to measure the effect of persuasion strategies on the engagement of each participant, considering both the different brain lobes that are activated, and the debate phases (the latter is the *within-subject*

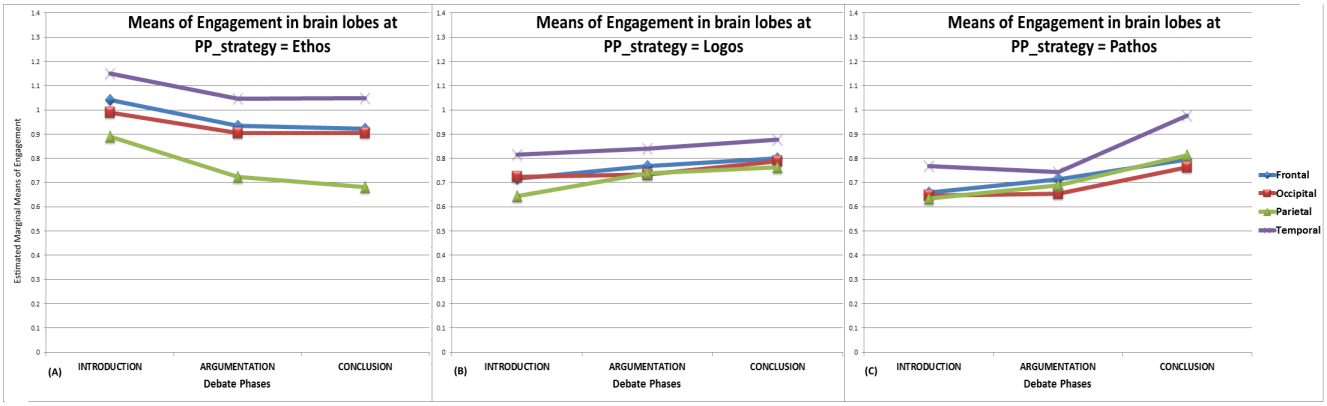


Figure 6.7: Estimated marginal means of engagements (y axis) in brain lobes by debates' phases (x axis) for the different persuasion strategies. Blue, red, green and violet lines correspond to the Frontal, Occipital, Parietal and Temporal brain lobes.

factor). As *between-subjects factors*, we consider the strategies and the brain lobes. Considering the resulting correlations among the strategy applied and the brain lobes activated in the participants in the different phases of the debate, we found $F(1.243, 30.683)$=4.495 and p=0.027. The factor *Deb_phases* has a significant effect on the participant's engagement. We also have a significant interaction of the factors $Deb\_phases * PP\_strategy$ with $F(2.486, 30.683)$=4.059 and p=0.012, meaning a significant effect on engagement[16]. The between-subject effects results show that there is a significant main effect of *PP_strategy* on the engagement, $F(2, 148)$=3.885, p=0.023.

For the Logos strategy, the most activated brain region is the parietal. Fig. 6.7-(B) shows that there is a significant difference between the INTRO and ARG phases for the parietal, which is the most activated lobe. By looking at the simple effect comparison, we found that the only significant mean difference is of parietal engagement between the ARG and INTRO phases with the Logos strategy (Mean difference= .115, $p = .019$). This result was unexpected, as we know that the frontal lobe is normally in charge of the planning, and rational decisions. By analyzing Logos arguments, we find that the PP used examples to justify her point of view, and imagination, residing in the parietal lobe, was triggered.

For the Ethos strategy, we have found that the parietal region was also activated. Looking at Fig. 6.7-(A), we see that the engagement in the parietal is high in the INTRO phase and decreases in the ARG phase. By looking at the simple effect comparison, we found that the only significant mean difference is of parietal engagement between the ARG and INTRO phases with the Ethos strategy (Mean difference = $-.174$, $p = .024$). For the CONC phase, the engagement remains similar to the ARG phase both with the Logos and Ethos strategies. Engagement is related to the resistance towards the persuader's arguments: the more there is a resistance, the more there is engagement. For the Ethos strategy, as the PP is assimilated to an expert, the engagement is decreasing in the ARG phase. Parietal lobes play a role in interpreting sensory information and orientation, meaning that the participant tries to establish new rules to take decisions. Recent studies discuss the correlation between this region and the process of decision making [172], and other studies have shown the role of right temporal-parietal junction for thinking about thoughts, e.g., people's belief, desires and emotions [284].

[16]Complete SPSS's results: `http://bit.ly/2nmbygV`.

For the Pathos strategy, the PP tried to induce empathy in participants. This resulted in the generation of strong emotions, and the circuit of emotions starts from the frontal to reach, through the cingulate Cortex, amygdala and hippocampus in the limbic system. The most important difference of engagement between the INTRO and ARG phases is indeed in the frontal lobe (see Fig. 6.7-(C)). In the simple effect analysis, the mean difference of the frontal engagement between INTRO and ARG with the Pathos strategy is the most important compared to the other brain lobes, even if it is not statistically significant (Mean diff.=0.61, $p = 0.332$).

H3 - Pathos persuasiveness. We hypothesize (H3) that the Pathos strategy impacts more than the other strategies in terms of persuasive power, and consequently it gathers more support towards the PP's arguments than the others. Table 6.4 reports about the changes of opinion of participants by comparing their initial opinion, and the final opinion after the debate. Since self-reporting is not predictive [292], the table reports also about participants who have changed their opinions but did not disclose this change in the questionnaire.

Debate	Strategy	PP position	P1	P2	P3	P4
DeathPenalty	Pathos	Con	Y	N	N	Y
Torture	Logos	Pro	N	Y	N	*Y*
Suicide	Ethos	Pro	N	N	N	*Y*
Profiling	Logos	Con	N	N	Y	*Y*
Nuclear	Logos	Pro	N	N	N	Y
Religion	Pathos	Con	N	N	*Y*	Y
Vaccines	Logos	Pro	N	N	N	N
GunRights	Ethos	Con	N	N	N	Y
Schools	Logos	Pro	N	N	*Y*	N
Organs	Pathos	Con	N	N	*Y*	*Y*

Table 6.4: Participants' changes of opinion. Y: an opinion change occurred; N: no change; *underlined*: change from neutral; *italic*: a change not reported by the participant (detected by comparing his initial and after-debate opinions).

To verify this hypothesis, we first need to normalize the number of attacks and supports for each debate wrt. the different strategies. Fig. 6.8 shows that the number of attacks and supports significantly changes depending on the strategy employed by the PP: Ethos is the strategy leading to the higher percentage of attacks in the argumentation, much more than the Logos and the Pathos strategies, while Pathos is the strategy leading to the higher percentage of supports wrt. the arguments proposed by the PP. Logos is in-between, as it is the most balanced strategy wrt. the percentage of attacks and supports. These results confirmed from the argumentation perspective what we already observed in H1 and H2: Pathos leads to the higher empathy leading to more supports than the other strategies. Note that these supports come even from those participants who do not agree with the PP, but they "cannot" attack the Pathos arguments she proposes, so they tend to agree on minor points related to the main topic. Ethos leads to more attacks than Logos: this can be explained by the fact that when an Ethos argument is proposed, the other participants do not evaluate the source as reliable, and tend to attack these arguments asking for evidences. Given that participants do not know each other, this behavior makes sense as authority is assessed by reputation and recommendation, and not only by claims.

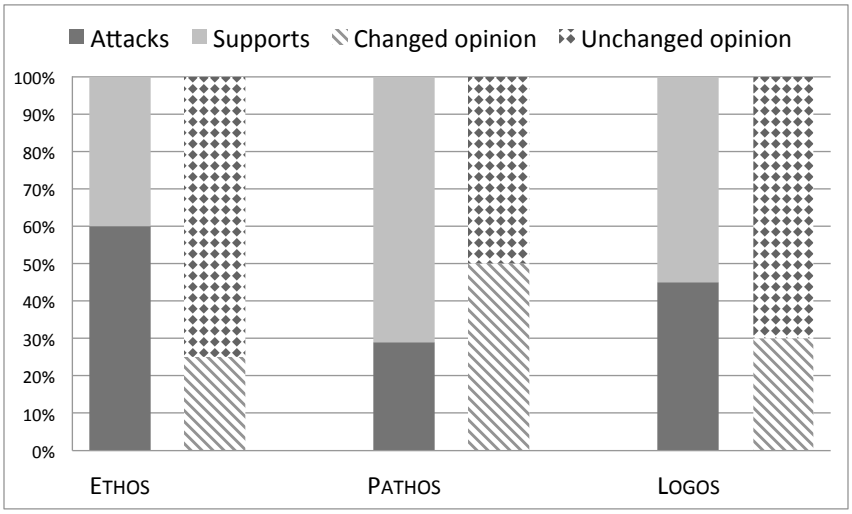


Figure 6.8: Percentage of attacks and supports for and against PP's arguments (1st columns), and percentage of participants with changed/unchanged opinion (2nd columns).

The validation of H3 is confirmed by analyzing the percentage of participants who changed/did not change their opinions wrt. the persuasion strategies (see Figure 6.8). On the one side, Pathos is the most effective strategy wrt. the percentage of participants who actually changed their mind after the debate, in line with the fact that participants tend to support Pathos arguments in the debate. On the other side, Logos and particularly Ethos are the less effective strategies with few participants persuaded by the PP.

6.6 Related Work

A first analysis of the experimental results presented in this chapter has been proposed in [39]. However, several aspects of the collected data were neglected in that work. In this extended version, the following issues have been tackled:

- *Personality traits, emotions and argumentation*: in [39], we did not consider in our analysis the Big Five inventory data we collected during the experiments. In this chapter, an additional hypothesis is formulated concerning the connection among participants' personality and emotions. The hypothesis has been then validated on the data collected from our experiments.

- *Opinions and emotions*: in [39], we did not consider the opinions of the participants with respect to the debated topics, their possible change during the debate, and the emotions. In this chapter, a fifth hypothesis is formulated and then validated on the collected data.

- *Correlations on single debates*: in [39], we considered correlations holding over the whole set of debates, i.e., over the whole set of collected data. However, we realized that some of the debates showed significant correlations that were not present in others, due to the involvement of different participants and to the interest of the participants in the debated topic. In this chapter, we have proposed an analysis of some of the more relevant debates held in our experimental sessions. These single debate

analysis considers also the workload index, computed at the data collection time but never discussed in the first version of the chapter [39].

- *Fourth layer of annotation*: in [39], three layers of annotation have been proposed over the collected textual data. In this chapter, we included a fourth annotation layer with the aim to highlight the change in the viewpoint of the participants during the debate.

In the literature, only few works deal with empirical experiments involving human participants to verify assumptions from argumentation theory. Cerutti et al. [96] propose an empirical experiment with humans in the argumentation theory area. However, the goal of this work is different from ours, since emotions are not considered and their aim is to show a correspondence between the acceptability of arguments by human subjects and the acceptability prescribed by the formal theory in argumentation. Rahwan and colleagues [267] study whether the meaning assigned to the notion of *reinstatement* in abstract argumentation theory is perceived in the same way by humans. They propose to the participants of the experiment a number of natural language texts where reinstatement holds, and then ask them to evaluate the arguments. Also in this case, the purpose of the work differs from ours, and emotions are not considered at all.

Emotions are considered, instead, by Nawwab et al. [234] that propose to couple the model of emotions introduced by Ortony and colleagues [243] in an argumentation-based decision making scenario. They show how emotions, e.g., gratitude and displeasure, impact on the practical reasoning mechanisms. A similar work has been proposed by Dalibon et al. [119] where emotions are exploited by agents to produce a line of reasoning according to the evolution of its own emotional state. Finally, Lloyd-Kelly and Wyner [209] propose emotional argumentation schemes to capture forms of reasoning involving emotions. All these works differ from our approach since they do not address an empirical evaluation of their models, and emotions are not detected from humans.

Several works in philosophy and linguistics have studied the link between emotions and natural argumentation, like [87, 155, 317]. These works analyze the connection of emotions and the different kind of argumentation that can be addressed. The difference with our approach is that they do not verify their theories empirically, on emotions extracted from people involved in an argumentation task. A particularly interesting case is that of the connection between persuasive argumentation and emotions, studied for instance by DeSteno and colleagues [124].

Concerning the empirical study of workload and emotional changes, [320] study pupillary response to detect workload and emotional changes performing an arithmetical task associated with pleasant/unpleasant images. The idea of the empirical study on workload and emotional changes is similar, even if the goal of the experiment is different, as our goal is connected to the argumentative process and not with arithmetical tasks performed by isolated participants.

Very few approaches in persuasive argumentation involve humans in the loop. Among them,[276] evaluate a methodology for human persuasion through argumentative dialogs, with human users. The huge difference wrt. [276] is that we do not analyze the argumentation style, but we capture the emotions and mental states directly on human participants through sensors. In [39], we studied the connections between emotions and argumentation, but we do not consider persuasion. In[38], we studied the correlation of the engagement index in brain hemispheres with the persuasion strategies. The difference with H2 is twofold: *i)* here, we provide a more fine grained analysis of the correlation of theengagement wrt. the four lobes instead of the left and right sides, *ii)* we concentrate on the correlation with the persuasion strategies, while in [38] we correlated with the neutral vs. opinionated (pro/con) stance of the participants. To the best of our knowledge, in neuroscience [80], no other work investigates the correlation between persuasive argumenta-

tion and mental states captured from users' brain through sensors. Usually, these factors are studied based on questionnaires with the participants.

6.7 Conclusion

In this chapter, we have presented an investigation into the links between the argumentation people use when they debate with each other, the emotions they feel during these debates, and their personality traits. We conducted an experiment aimed at verifying our hypotheses about the correlation between the positive/negative emotions emerging when positive/negative relations among the arguments are put forward in the debate, and the correlation between the personality traits of the debaters and their opinions on the debated topics, and the emotions felt during the debate interactions. The results suggest that there exist trends that can be extracted from emotion analysis. Moreover, we also provide the first annotated dataset and gold standard to compare and analyze emotion detection in an argumentation session.

The take-home message of this chapter is twofold: first, high engagement is correlated with negative emotions showing that participants are mentally involved in producing arguments to rebut those which are not in line with their viewpoint, and second, neuroticism and conscientiousness have both a negative impact on the debaters' brain indexes ending up into a reduced mental engagement index and an increased cognitive load. Finally, the surprise emotion is shown by extroverted debaters.

The main contributions concerning persuasion in human argumentation are: *i)* the first field experiment to study the correlation of persuasion strategies, argumentation and emotions using EEG headsets and cameras, *ii)* an annotated dataset of arguments characterized by a persuasion strategy, and *iii)* the first steps towards the definition of human-like empathic argumentative agents.

The analysis of the results allowed us to highlight some drawbacks of our experimental setting to be addressed: *i)* more fine grained persuasion strategies may be considered, as these categories are highly general and sometimes difficult to be evaluated; *ii)* the strategies adopted by the other participants should be taken into account to expand the scenario (here, to overcome this issue, we consider them as random and we focus on the punctual reactions of the participants to PP's arguments); *iii)* the binary variable (pro/con) expressing the stance of the participants wrt. the debated issue may not fully capture the effect of a strategy, so allowing the expression of degrees of pro/con could be preferable.

Several lines of research have to be considered as future work. First, we aim to study how emotions persistence influences the attitude of the debates: this kind of experiment has to be repeated a number of times in order to verify whether positive/negative emotions before the debate influence new interactions. Second, we plan to add a further step, namely to study how sentiment analysis methods developed in Computational Linguistics are able to automatically detect the polarity of the arguments proposed by the debaters, and how they are correlated with the detected emotions. More precisely, the annotated dataset we published provides a valuable resource to improve the performances of sentiment analysis systems allowing them to learn about the correlation among the relations among the arguments and the emotions aligned with the arguments. Moreover, we plan to study emotions propagation among the debaters, and to verify whether an emotion can be seen as a predictor of the solidity of an argument, e.g., if I write an argument when I am angry I may make wrong judgments. Finally, argumentation theory has often been proposed as a technique for supporting *critical thinking*, thus studying the relation of these philosophical theories with emotions and personality traits of the actors involved in the argumentation is a further step to investigate.

Chapter 7

Conclusion and perspectives

This document provided an overview of my research activity from 2010 to 2018. It has taken place within various local, national and international projects. My research areas are Artificial Intelligence, Knowledge Representation and Reasoning, Argumentation Theory and Normative Reasoning, with the general aim of supporting machine decision making by providing formal models able to explain the reasons behind the decisions, whether the sources of information are trustable or not, and how emotions impact the deliberation process. During this 8-year period, I addressed the following three general research questions:

1. how to define argumentation models able to support decision making and to be used to justify the deliberation process?

2. how to mine arguments from documents in natural language and what is their relation with the emotions and mental states of their proposers?

3. how to mine and reason about normative information?

My contributions have been published in main international conferences and journals of my research communities. They mainly deal with:

1. Computational models of argument, focusing on the definition of formal models of argument with the goal of supporting deliberation and explanation, taking into account external components like trust, emotions and norms;

2. Natural models of argument, focusing on the definition of empirical methods for detecting argumentative structures and predicting the relations among them from natural language texts, considering application scenarios like social media (e.g., Twitter, Wikipedia, Debatepedia), medical trials, and political debates;

3. Norm mining and normative reasoning, focusing on the definition of both empirical and formal methods to extract and reason on norms and rights information to define the next generation of legal informatics systems.

In this document I have made the decision of presenting a number of my contributions that show my unconventional journey in the field of computational models of argument, starting from a formal approach to define modularity and decomposability in abstract argumentation frameworks to the modeling of the notion

of trust using argumentation theory, to the definition of mining algorithms for natural language argumentation, to the study of emotions in human argumentation.

In the continuation of my ongoing research work, and in line with the objectives of the Wimmics team, I will keep tackling the three general research questions discussed in this document, in line with the recents trends in Artificial Intelligence. Since the field's early years, Artificial Intelligence (AI) has the goal to understand the principles governing intelligent behavior and to encode such principles in so called *intelligent machines*. In the latest years, progress in AI seems to be accelerating, given the recent results Machine Learning, Natural Language Processing and Computer Vision, leading to important investments in AI from main information technologies companies like Google, and IBM. These companies see in this field new potential markets. However, together with the increasing popularity of AI and the expectations on it, new concerns are now rising around the development of super-intelligent machines. Actually, these concerns are not new, as Turing pointed out in 1951 *"If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled. ... [T]his new danger ... is certainly something which can give us anxiety."*. More recently, popular public figures like Elon Musk[1] and Stephen Hawking also underlined the negative impact that super-intelligent machines may have on society, even fostering apocalyptic scenarios where machines take the control of the human society. It must be said, however, that even if many *precise* intelligent tasks can be performed autonomously by machines, still the emulation of human behavior is far [226, 65, 287, 282, 281]. Important human-like intelligence features like the ability to conceive new actions, and the capability of being conscious of our own decisions and their consequences are unachievable by current intelligent machines, as Searle pointed out after the victory of IBM Watson on the TV show "Jeopardy!" *"Watson did not understand the questions, nor its answers, not that some of its answers were right and some wrong, not that it was playing a game, nor that it won."* [287].

The panorama offered to AI researchers is thus the following: on the one hand, many of the goals of the field's early years are becoming reality, but on the other hand, the question to be asked to ourselves is whether we still actually want human-level AI, considering the risks that are in front of us [2]. It must be stressed at this point that AI can be enormously beneficial for human flourishing, but we need to take care about the design of AI in order to reach a so-called good AI hybrid society. In this society, mixed teams of intelligent machines and humans jointly realize human values and promote the public good. Using Russell's words, *"These problems require a change in the definition of AI itself, from a field concerned with pure intelligence, independent of the objective, to a field concerned with systems that are provably beneficial for humans."*, and Wooldridge's ones *"The technology is right here, right now. It's just a matter of rolling it out and the big social and legal issues that go with that."*.[3] The definition of a "good AI society" [65] or "beneficial AI" [281] may sound like an utopia, but it is not the case. It is a matter of ensuring that these machines, with an arbitrary degree of autonomy and intelligence, remain under human control, and doing so is possible only if we start from now to design them in a way such as their goal is the preservation of human values and dignity. My future research goes in this direction, and sets to tackle the challenge of designing intelligent machines towards a good AI hybrid society.

My long-term program is to design intelligent machines with the capability to form machine-human teams acting in a good AI hybrid society. In a short- and mid-term perspective, my goal is to answer these

[1] `https://goo.gl/RFGfzm`

[2] We refer the reader to Russell [281], who discussed the danger in dismissing the arguments against AI with too much levity.

[3] The root of this trend may be found in the 60s when there was the trend of Intelligence Amplification [10] and Intelligence Augmentation [142] that addressed this idea by considering from the beginning that the goal is to help humans.

precise research questions:

- how to design intelligent machines able to explain and justify to humans their decisions?
- how to design intelligent machines grounding their decisions on moral values?
- how to design intelligent machines employing sentimental values in their deliberation process?

Explanations and justifications for intelligent machine deliberation. Machine accountability is strongly linked to explanation and justification, as included in the new European General Data Protection Regulation, and needs to be grounded in moral and social concepts, including moral and emotional values. Every intelligent machine acting for a good AI hybrid society should operate within a moral and social framework, in explainable and justified ways. It goes without saying that they must operate within the bounds of the law, including, for example, the legal requirements associated with the handling of the user data acquired and collected to improve predictions, suggestions and response times. The full impact of these legal requirements may soon impact the technical requirements of the next generation of intelligent machines, requiring new types of collaboration between lawyers and computer scientists.

Only very few approaches have tackled the problem of explaining the decisions taken by intelligent machines. For instance, Lei et al. [196] propose a new way to train neural networks so that they provide not only predictions and classifications but rationales for their decisions. The need for explainability is even more pronounced with recent advances in neural models. Some efforts in this area include analyzing and visualizing state activation [169, 186, 199], and linking word vectors to semantic lexicons or word properties [145, 168]. Also attention based models have been successfully applied to many Natural Language Processing problems, improving visualization and interpretability [280, 273, 169].

Despite these approaches, the explanation and justification of the decisions of intelligent machines are still far from being achieved. The main open challenge is that these approaches do not return explanations understandable by humans, as they are more likely to be used for justifying the output to other domain experts. What I target, instead, is the definition of appropriate methods to explain machine decisions to whoever human, in a way humans can understand and more importantly interact. This is a key point of the explanation methods on which I focus: humans need to understand machines and their decisions through the arguments they propose to explain their course of action, and this process is an interactive process so that humans can ask for more details about certain arguments, they can ask for more support to an argument, and finally they may at their turn explain why a certain course of action is / is not compliant with the human values and/or the social welfare. This kind of explanation is what adults address to children to let them learn new values. I argue that the same approach should be applied to machines as our final objective is to make them behave in order to achieve human values.

To achieve this goal, we will rely on argumentation theory and on recent advances in processing natural language arguments (i.e., argument mining). My first goal is to enhance transparency in the interactions among intelligent machines and humans. The methodology to tackle this goal combines argument mining, and NLP methods in general, (to extract arguments from textual information sources, and as a further step to analyse written or oral speech and to generate natural language arguments), KRR formalisms like argumentation theory but also case-based reasoning and decision making models (to reason over the mined information), and semantic knowledge graphs (to provide machines with background knowledge). In order to allow intelligent machines to build their own arguments explaining their actions or claims, we will make them mine the biggest information source ever: the Web. This is a recursive procedure such that the interactions of the intelligent machines with humans will trigger mining further arguments to support specific

aspects of the deliberation. The output will consist of constrained Natural Language explanations, auditable by humans. The evaluation will be empirical (precision, recall, F1) for the argument mining part, together with a user evaluation to assess the human satisfaction about the returned explanations. The output will be enriched also by considering moral and sentimental values in the justification process. This future work is the natural continuation of my current research activity about argumentation theory and argument mining. The need of transparency will be addressed in the mid-term perspective concerning the scenarios of clinical trials (transparency and explanation of clinical decisions) and political debates (transparency and fallacies detection in political speeches and debates).

Ethics and responsibility for intelligent machine deliberation. Intelligent machine ethics is about understanding, developing and evaluating ethical agency and reasoning abilities as part of the behavior of AI systems (such as intelligent machines and robots). Even though intelligent machines are increasingly able to take decisions and perform actions that have a moral impact, they are still artifacts and therefore they are neither ethically nor legally responsible. Human beings should remain the moral agent. We can delegate control to purely synthetic intelligent machines without delegating responsibility or liability to them. With the term machine ethics, we refer to the computational and theoretical methods and tools that support the representation, evaluation, verification, and transparency of ethical deliberation by machines with the aim of supporting human values and human dignity on shared tasks with those machines. That is, machine ethics concerns the methods, algorithms, and tools needed to endow intelligent machines with the capability to reason about the ethical aspects of their decisions, and the ethically informed design guidelines for developing intelligent machines whose behavior is guaranteed to remain within acceptable ethical constraints. Research is needed to understand what suitable constraints on system behavior are, and to elicit desiderata on the representation and use of moral values by intelligent machines. One of my objectives is to identify appropriate methods for eliciting and representing ethical and emotional requirements, and suitable machine deliberation architectures for explicit reasoning in terms of moral and social features as one mechanism for handling moral dilemmas [56] and providing explanations of behavior through argumentation.

Some formal representation approaches to ethical behavior of intelligent machines address these issues but they mainly focus either on modeling moral reasoning [56] as a direct translation of some well-known moral theory, on modeling moral machines in a general way [211], or on designing an ethical intelligent machine architecture. Such architectures include implicit ethical architectures [7] which design the machines's behavior (either by implementing or learning) for each situation in order to avoid potential unethical behaviors, or cognitive ethical architectures [108, 109] consisting of full explicit representations of each component of the machine, from the classical beliefs (information on the environment and other machines), desires (goals of the intelligent machine) and intentions (the chosen actions), to heuristics and emotional machinery. Even though these approaches successfully address some problems [31, 4, 126, 110], the definition of general frameworks to model computational morality in AI is still a major and urgent open challenge. Furthermore, these approaches do not clearly take into account the collective and distributed dimension of the interaction.

I am also involved in the submitted COST Action proposal titled "Responsible Artificial Intelligence". This project aims at building a network of researchers from different disciplines such as computer science, law, philosophy, linguistics, and psychology. The project is about human responsibility for the development of intelligent systems along fundamental human principles and values, to ensure human flourishing and wellbeing in a sustainable world. I was also involved in the organization of the Thematic Day about "Ethics by design" organized in co-location with PRIMA-2017 conference.

My second goal is to define design principles for intelligent machines taking into account moral val-

ues in their deliberation process. Such principles will be implemented in moral-based reward functions to provide machine deliberation with the ethical dimension. I will consider decision-making to be performed through a collective deliberative procedure, i.e., a *moral dialogue*. The methodology to tackle this ambitious goal combines argumentation theory and legal reasoning to model moral dialogues. The role of norms to characterize moral autonomy has been differently, but successfully proposed by (competing) moral views like Kantianism and rule utilitarianism [185]. Drawing from these traditions, I will formally explore the following intuition: autonomous machines take decisions about the moral theory governing their society, and elicit new theories that would improve the welfare. Game theoretic approaches to compute reward functions will be enhanced with moral values, and argumentation-based conflict detection mechanisms will be included. The evaluation will be formal. The principles will be validated by the experts involved in the RAI COST Action and the MIREL partners. This future work is the natural continuation of my research about normative reasoning, in particular my current research activity connecting argumentation theory and normative reasoning. Another interesting application of these formal models is to make autonomous agents more resilient to attacks like the ones that targeted the Microsoft Agent to make it return racist/xenophobic/... arguments. The idea would be to use moral reasoning to protect the agent from bad influences and test it with replays or synthetic attacks.

Emotional values for intelligent machine deliberation. Human-level AI and human-AI teams imply the fact that the AI acting in this hybrid society actually learns human values. A very important part of these values deals with emotions and empathy, so intelligent machines for a good AI hybrid society cannot underestimate the importance of sentimental values for human beings. As these intelligent machines need to learn from human behavior what are the sentimental values, how to express them, and more importantly, how do they influence rational reasoning, a deep analysis of how emotional values appear in human beings is required. I also target the goal of first addressing field experiments to analyze how human emotions influence reasoning, and more specifically, how they influence explanations and arguments acceptance, and second, using the obtained results to design emphatic intelligent machines able to cope with human emotional values.

Some formal representation approaches to representing emotional values in intelligent machines have been proposed, but they mainly focus on proposing formal frameworks including emotions in the reasoning process [234, 119, 209]. None of these works has addressed first an empirical evaluation to analyze how this kind of reasoning actually holds for human beings. Moreover, their goal is not to define a framework such that emotional values elicit the goals and actions of intelligent machines.

My third set of goals is *i)* to define the principles to design intelligent machines taking into account emotional values in their deliberation process, and *ii)* to formally define reward functions ensuring that such emotional values play a role in decision-making. I will study the impact of emotions in deliberation dialogues by addressing a new field experiment with humans. I will use the FaceReader software to recognize basic emotions (i.e., *happy*, *sad*, *angry*, *surprised*, *scared* and *disgusted*), and I will also consider mental states (captured by using Emotiv Epoc EEG headsets), that is, mental engagement (i.e., the level of attention in performing a task [260, 146]), and workload (i.e., the quantity of information processing required to perform a task [249]). The whole equipment, i.e., EEG headsets and FaceReader, will be provided by the CoCoLab platform of the Université Côte d'Azur-CNRS. Emotion-based reward functions for machine deliberation will be defined based on the results of such field experiment. The evaluation is empirical for what concerns the data collected from the field experiment (ANOVA, MANOVA), whilst it is formal for the emotional reward functions and the resulting framework. This future work is the natural continuation of my research line about argumentation and emotions.

To summarize, my future research perspectives shared the common goal to define and develop AI solu-

tions to support human values and and promote the public good. To tackle this challenging and wide goal, I plan to investigate three main research directions: *i)* explaining and justifying machine decisions through natural language arguments and reasoning formalisms, *ii)* defining formal models for ethical and responsible machine deliberation, and *iii)* defining machine deliberation models based on human emotional values.

Bibliography

[1] Ehud Aharoni et al. “A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics”. In: *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, 2014, pp. 64–68.

[2] Leila Amgoud and Philippe Besnard. “Logical limits of abstract argumentation frameworks”. In: *Journal of Applied Non-Classical Logics* 23.3 (2013), pp. 229–267. DOI: 10.1080/11663081.2013.830381. URL: https://doi.org/10.1080/11663081.2013.830381.

[3] Leila Amgoud and Henri Prade. “Reaching Agreement Through Argumentation: A Possibilistic Approach”. In: *KR*. Ed. by Didier Dubois, Christopher A. Welty, and Mary-Anne Williams. AAAI Press, 2004, pp. 175–182. ISBN: 1-57735-199-1.

[4] Dario Amodei et al. “Concrete Problems in AI Safety”. In: *CoRR* abs/1606.06565 (2016). URL: http://arxiv.org/abs/1606.06565.

[5] Ion Androutsopoulos and Prodromos Malakasiotis. “A survey of paraphrasing and textual entailment methods”. In: *J. Artif. Int. Res.* 38.1 (2010), pp. 135–187. ISSN: 1076-9757. URL: http://dl.acm.org/citation.cfm?id=1892211.1892215.

[6] Alessio Palmero Aprosio, Claudio Giuliano, and Alberto Lavelli. “Automatic Expansion of DBpedia Exploiting Wikipedia Cross-Language Information”. In: *ESWC*. Ed. by Philipp Cimiano et al. Vol. 7882. Lecture Notes in Computer Science. Springer, 2013, pp. 397–411. ISBN: 978-3-642-38287-1. DOI: 10.1007/978-3-642-38288-8. URL: https://doi.org/10.1007/978-3-642-38288-8.

[7] Ronald C. Arkin. “Ethical robots in warfare”. In: *IEEE Technol. Soc. Mag.* 28.1 (2009), pp. 30–33. DOI: 10.1109/MTS.2009.931858. URL: https://doi.org/10.1109/MTS.2009.931858.

[8] Ivon Arroyo et al. “Emotion Sensors Go To School”. In: *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2009, pp. 17–24. ISBN: 978-1-60750-028-5. URL: http://dl.acm.org/citation.cfm?id=1659450.1659458.

[9] R. Artstein. “Inter-annotator Agreement”. In: *Handbook of Linguistic Annotation*. Ed. by Nancy Ide and James Pustejovsky. Dordrecht: Springer Netherlands, 2017, pp. 297–313. ISBN: 978-94-024-0881-2. DOI: 10.1007/978-94-024-0881-2_11. URL: https://doi.org/10.1007/978-94-024-0881-2_11.

[10] W.R. Ashby. *An Introduction to Cybernetics*. 1963. URL: https://books.google.fr/books?id=-V4YGwHOUs0C.

[11] K. Atkinson. "Introduction to special issue on modelling Popov v. Hayashi". In: *Artificial Intelligence and Law* 20.1 (2012), pp. 1–14.

[12] R. Bar-Haim et al. "Stance Classification of Context-Dependent Claims". In: *EACL*. 2017.

[13] P. Baroni, M. Caminada, and M. Giacomin. "An introduction to argumentation semantics". In: *The Knowledge Engineering Review* 26.4 (2011), pp. 365–410.

[14] P. Baroni, P. E. Dunne, and M. Giacomin. "On the resolution-based family of abstract argumentation semantics and its grounded instance". In: *Artificial Intelligence* 175.3-4 (2011), pp. 791–813.

[15] P. Baroni and M. Giacomin. "On principle-based evaluation of extension-based argumentation semantics". In: *Artificial Intelligence (Special issue on Argumentation in A.I.)* 171.10/15 (2007), pp. 675–700.

[16] P. Baroni and M. Giacomin. "Skepticism relations for comparing argumentation semantics". In: *International Journal of Approximate Reasoning* 50.6 (2009), pp. 854–866.

[17] P. Baroni and M. Giacomin. "Solving Semantic Problems with Odd-Length Cycles in Argumentation". In: *Proc. of the 7th European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2003)*. 2003, pp. 440–451.

[18] P. Baroni, M. Giacomin, and B. Liao. "On topology-related properties of abstract argumentation semantics. A correction and extension to *Dynamics of argumentation systems: A division-based method*". In: *Artificial Intelligence* 212 (2014), pp. 104–115.

[19] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. "An introduction to argumentation semantics". In: *Knowledge Eng. Review* 26.4 (2011), pp. 365–410. DOI: 10.1017/S0269888911000166. URL: http://dx.doi.org/10.1017/S0269888911000166.

[20] Pietro Baroni et al. "On the Input/Output behavior of argumentation frameworks". In: *Artif. Intell.* 217 (2014), pp. 144–197. DOI: 10.1016/j.artint.2014.08.004. URL: https://doi.org/10.1016/j.artint.2014.08.004.

[21] P. Baroni et al. "AFRA: Argumentation framework with recursive attacks". In: *Int. J. Approx. Reasoning* 52.1 (2011), pp. 19–37.

[22] P. Baroni et al. "Encompassing Attacks to Attacks in Abstract Argumentation Frameworks". In: *Proc. of the 10th European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2009)*. 2009, pp. 83–94.

[23] P. Baroni et al. "On Input/Output Argumentation Frameworks". In: *Proc. of the 4th Int. Conf. on Computational Models of Argument (COMMA 2012)*. 2012, pp. 358–365.

[24] R. Barzilay and K.R. McKeown. "Sentence fusion for multidocument news summarization." In: *Computational Linguistics* 31(3) (2005), pp. 297–327.

[25] R. Baumann. "Normal and strong expansion equivalence for argumentation frameworks". In: *Artificial Intelligence* 193 (2012), pp. 18–44.

[26] R. Baumann. "What Does it Take to Enforce an Argument? Minimal Change in Abstract Argumentation". In: *Proc. of the 20th European Conf. on Artificial Intelligence (ECAI 2012)*. 2012, pp. 127–132.

[27] R. Baumann and G. Brewka. "Analyzing the Equivalence Zoo in Abstract Argumentation". In: *Proc. of the 14th Int. Workshop on Computational Logic in Multi-Agent Systems (CLIMA XIV)*. 2013, pp. 18–33.

[28] R. Baumann, G. Brewka, and R. Wong. "Splitting Argumentation Frameworks: An Empirical Evaluation". In: *Theory and Applications of Formal Argumentation - First Int. Workshop (TAFA 2011). Revised Selected Papers*. Vol. 7132. Lecture Notes in Computer Science. Springer, 2011, pp. 17–31.

[29] Ringo Baumann. "Splitting an Argumentation Framework". In: *Proc. of LPNMR 2011 11th Int. Conf. on Logic Programming and Nonmonotonic Reasoning*. 2011, pp. 40–53.

[30] R. Baumann et al. "Parameterized Splitting: A Simple Modification-Based Approach". In: *Correct Reasoning - Essays on Logic-Based AI in Honour of Vladimir Lifschitz*. Ed. by E. Erdem et al. Vol. 7265. Lecture Notes in Computer Science. Springer, 2012, pp. 57–71.

[31] Aline Belloni et al. "Dealing with Ethical Conflicts in Autonomous Agents and Multi-Agent Systems". In: *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January 25, 2015*. Ed. by Toby Walsh. Vol. WS-15-02. AAAI Workshops. AAAI Press, 2015. ISBN: 978-1-57735-713-1. URL: http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10109.

[32] Amel Ben Othmane et al. "A Multi-context BDI Recommender System: from Theory to Simulation". In: *Web Intelligence*. 2016 IEEE/WIC/ACM International Conference on Web Intelligence. Omaha, United States, Oct. 2016. DOI: 10.1109/WI.2016.0104. URL: https://hal.archives-ouvertes.fr/hal-01400997.

[33] Amel Ben Othmane et al. "A Multi-context Framework for Modeling an Agent-Based Recommender System". In: *8th International Conference on Agents and Artificial Intelligence (ICAART2016)*. 2016. URL: https://hal.inria.fr/hal-01239884.

[34] Amel Ben Othmane et al. "Towards a Spatio-Temporal Agent-Based Recommender System". In: *16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2017)*. Ed. by Kate Larson et al. Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8–12, 2017. ACM. São Paulo, Brazil: ACM, May 2017. URL: https://hal.archives-ouvertes.fr/hal-01531169.

[35] Trevor J. M. Bench-Capon, D. Lowes, and A. M. McEnery. "Argument-based explanation of logic programs". In: *Knowl.-Based Syst.* 4.3 (1991), pp. 177–183.

[36] Trevor J. M. Bench-Capon and Giovanni Sartor. "Theory based explanation of case law domains". In: *ICAIL*. 2001, pp. 12–21.

[37] S. Benferhat and S. Kaci. "Logical representation and fusion of prioritized information based on guaranteed possibility measures: application to the distance-based merging of classical bases". In: *Artif. Intell.* 148.1-2 (2003), pp. 291–333. ISSN: 0004-3702.

[38] Mohamed S. Benlamine et al. "Persuasive Argumentation and Emotions: An Empirical Evaluation with Users". In: *Proc. of HCI 2017*. 2017, pp. 659–671.

[39] Sahbi Benlamine et al. “Emotions in Argumentation: an Empirical Evaluation”. In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. Ed. by Qiang Yang and Michael Wooldridge. AAAI Press, 2015, pp. 156–163. ISBN: 978-1-57735-738-4. URL: http://ijcai.org/papers15/Abstracts/IJCAI15-029.html.

[40] James Bergstra and Yoshua Bengio. “Random Search for Hyper-parameter Optimization”. In: *J. Mach. Learn. Res.* 13.1 (2012), pp. 281–305. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2503308.2188395.

[41] Philippe Besnard and Anthony Hunter. “A logic-based theory of deductive arguments”. In: *Artif. Intell.* 128.1-2 (2001), pp. 203–235.

[42] Philippe Besnard and Anthony Hunter. *Elements of Argumentation*. MIT Press, 2008. ISBN: 978-0-262-02643-7. URL: http://mitpress.mit.edu/books/elements-argumentation.

[43] Floris Bex and Douglas Walton. “Burdens and Standards of Proof for Inference to the Best Explanation”. In: *Legal Knowledge and Information Systems - JURIX 2010: The Twenty-Third Annual Conference on Legal Knowledge and Information Systems, Liverpool, UK, 16-17 December 2010*. Ed. by Radboud Winkels. Vol. 223. Frontiers in Artificial Intelligence and Applications. IOS Press, 2010, pp. 37–46. ISBN: 978-1-60750-681-2. DOI: 10.3233/978-1-60750-682-9-37. URL: http://dx.doi.org/10.3233/978-1-60750-682-9-37.

[44] Or Biran and Owen Rambow. “Identifying Justifications in Written Dialogs by Classifying Text as Argumentative”. In: *Int. J. Semantic Computing* 5.4 (2011), pp. 363–381. DOI: 10.1142/S1793351X11001328. URL: http://dx.doi.org/10.1142/S1793351X11001328.

[45] Stefano Bistarelli, Fabio Rossi, and Francesco Santini. “A First Comparison of Abstract Argumentation Reasoning-Tools”. In: *ECAI 2014 - 21st European Conference on Artificial Intelligence*. 2014, pp. 969–970.

[46] Stefano Bistarelli, Fabio Rossi, and Francesco Santini. “A First Comparison of Abstract Argumentation Systems: A Computational Perspective”. In: *Proceedings of the 28th Italian Conference on Computational Logic*. 2013, pp. 241–245.

[47] Stefano Bistarelli, Fabio Rossi, and Francesco Santini. “Benchmarking Hard Problems in Random Abstract AFs: The Stable Semantics”. In: *Computational Models of Argument - Proceedings of COMMA 2014*. 2014, pp. 153–160.

[48] P. Blackburn et al. “Inference and computational semantics.” In: *Studies in Linguistics and Philosophy, Computing Meaning* 77(2) (2001), 11Ð28.

[49] David M. Blei and John D. Lafferty. “Correlated topic models”. In: *In Proceedings of the 23rd International Conference on Machine Learning*. MIT Press, 2006, pp. 113–120.

[50] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. “Latent Dirichlet Allocation”. In: *J. Mach. Learn. Res.* 3 (2003), pp. 993–1022. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944937.

[51] Joshua E. Blumenstock. “Size Matters: Word Count As a Measure of Quality on Wikipedia”. In: *Proceedings of the 17th International Conference on World Wide Web*. WWW ’08. Beijing, China: ACM, 2008, pp. 1095–1096. ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367673. URL: http://doi.acm.org/10.1145/1367497.1367673.

[52] Alexander Bochman. "A causal approach to nonmonotonic reasoning". In: *Artif. Intell.* 160.1-2 (2004), pp. 105–143.

[53] Guido Boella, Leendert van der Torre, and Serena Villata. "On the Acceptability of Meta-arguments". In: *Procs. of the 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT-2009)*. IEEE, 2009, pp. 259–262.

[54] Guido Boella et al. "Meta-Argumentation Modelling I: Methodology and Techniques". In: *Studia Logica* 93.2-3 (2009), pp. 297–355.

[55] Guido Boella et al. "Support in Abstract Argumentation". In: *COMMA*. Ed. by Pietro Baroni et al. Vol. 216. Frontiers in Artificial Intelligence and Applications. IOS Press, 2010, pp. 111–122. ISBN: 978-1-60750-618-8.

[56] Vincent Bonnemains, Saurel Claire, and Catherine Tessier. "How Ethical Frameworks Answer to Ethical Dilemmas: Towards a Formal Model". In: *Proceedings of the 1st Workshop on Ethics in the Design of Intelligent Agents, The Hague, Holland, August 30, 2016.* Ed. by Grégory Bonnet et al. Vol. 1668. CEUR Workshop Proceedings. CEUR-WS.org, 2016, pp. 44–51. URL: http://ceur-ws.org/Vol-1668/paper8.pdf.

[57] R. Booth et al. "Conditional Acceptance Functions". In: *Proc. of the 4th Int. Conf. on Computational Models of Argument (COMMA 2012)*. 2012, pp. 470–477.

[58] J. Bos and K. Markert. "When Logical Inference helps determining textual entailment (and when it doesn't)". In: *Proc. of the 2nd PASCAL Workshop on Recognizing Textual Entailment*. 2006.

[59] T. Bosc, E. Cabrio, and S. Villata. "Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media". In: *COMMA 2016*. 2016, pp. 21–32.

[60] Tom Bosc, Elena Cabrio, and Serena Villata. "DART: a Dataset of Arguments and their Relations on Twitter". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.* Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA), 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/611.html.

[61] Tom Bosc, Elena Cabrio, and Serena Villata. "DART: a Dataset of Arguments and their Relations on Twitter (accepted for publication)". In: *Proceeding of LREC 2016*. 2016.

[62] G. Brewka and S. Woltran. "Abstract Dialectical Frameworks". In: *Proc. of the 12th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2010)*. 2010, pp. 102–111.

[63] Volha Bryl and Christian Bizer. "Learning conflict resolution strategies for cross-language Wikipedia data fusion". In: *WWW (Companion Volume)*. 2014, pp. 1129–1134.

[64] Katarzyna Budzynska and Chris Reed. *Whence inference*. Tech. rep. University of Dundee, 2011.

[65] Cath C. et al. "Artificial Intelligence and the 'Good Society': the US, EU, and UK approach." In: *Sci Eng Ethics.* (2017).

[66] E. Cabrio, B. Magnini, and A. Ivanova. "Extracting Context-Rich Entailment Rules from Wikipedia Revision History". In: *The People's Web Meets NLP Workshop*. 2012.

[67] E. Cabrio and S. Villata. "Natural Language Arguments: A Combined Approach". In: *European Conference on Artificial Intelligence (ECAI)*. 2012, pp. 205–210.

[68] Elena Cabrio, Alessio Palmero Aprosio, and Serena Villata. "Reconciling Information in DBpedia through a Question Answering System". In: *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014*. Ed. by Matthew Horridge, Marco Rospocher, and Jacco van Ossenbruggen. Vol. 1272. CEUR Workshop Proceedings. CEUR-WS.org, 2014, pp. 49–52. URL: http://ceur-ws.org/Vol-1272/paper_44.pdf.

[69] Elena Cabrio, Julien Cojan, and Fabien Gandon. "Mind the cultural gap: bridging language specific DBpedia chapters for Question Answering". In: *to appear in Towards the Multilingual Semantic Web*. Ed. by Philipp Cimiano and Paul Buitelaar. Springer Verlag, 2014.

[70] Elena Cabrio and Serena Villata. "A natural language bipolar argumentation approach to support users in online debate interactions". In: *Argument & Computation* 4.3 (2013), pp. 209–230. DOI: 10.1080/19462166.2013.862303. URL: http://dx.doi.org/10.1080/19462166.2013.862303.

[71] Elena Cabrio and Serena Villata. "Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions". In: *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*. The Association for Computer Linguistics, 2012, pp. 208–212. ISBN: 978-1-937284-25-1. URL: http://www.aclweb.org/anthology/P12-2041.

[72] Elena Cabrio and Serena Villata. "Natural Language Arguments: A Combined Approach". In: *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012*. Ed. by Luc De Raedt et al. Vol. 242. Frontiers in Artificial Intelligence and Applications. IOS Press, 2012, pp. 205–210. ISBN: 978-1-61499-097-0. DOI: 10.3233/978-1-61499-098-7-205. URL: https://doi.org/10.3233/978-1-61499-098-7-205.

[73] Elena Cabrio and Serena Villata. "NoDE: A Benchmark of Natural Language Arguments". In: *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*. Ed. by Simon Parsons et al. Vol. 266. Frontiers in Artificial Intelligence and Applications. IOS Press, 2014, pp. 449–450. ISBN: 978-1-61499-435-0. DOI: 10.3233/978-1-61499-436-7-449. URL: http://dx.doi.org/10.3233/978-1-61499-436-7-449.

[74] Elena Cabrio, Serena Villata, and Alessio Palmero Aprosio. "A RADAR for information reconciliation in Question Answering systems over Linked Data". In: *Semantic Web* 8.4 (2017), pp. 601–617. DOI: 10.3233/SW-160245. URL: https://doi.org/10.3233/SW-160245.

[75] Elena Cabrio, Serena Villata, and Fabien Gandon. "A Support Framework for Argumentative Discussions Management in the Web". In: *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*. Ed. by Philipp Cimiano et al. Vol. 7882. Lecture Notes in Computer Science. Springer, 2013, pp. 412–426. ISBN: 978-3-642-38287-1. DOI: 10.1007/978-3-642-38288-8_28. URL: https://doi.org/10.1007/978-3-642-38288-8_28.

[76] Elena Cabrio, Serena Villata, and Fabien Gandon. "Classifying Inconsistencies in DBpedia Language Specific Chapters". In: *Procs of LREC-2014*. 2014.

[77] Elena Cabrio et al. “Argumentation-based Inconsistencies Detection for Question-Answering over DBpedia”. In: *NLP-DBPEDIA@ISWC*. 2013.

[78] Elena Cabrio et al. “Hunting for Inconsistencies in Multilingual DBpedia with QAKiS”. In: *Proceedings of the ISWC 2013 Posters & Demonstrations Track, Sydney, Australia, October 23, 2013*. Ed. by Eva Blomqvist and Tudor Groza. Vol. 1035. CEUR Workshop Proceedings. CEUR-WS.org, 2013, pp. 69–72. URL: http://ceur-ws.org/Vol-1035/iswc2013_demo_18.pdf.

[79] Elena Cabrio et al. “QAKiS: an Open Domain QA System based on Relational Patterns”. In: *Procs of ISWC 2012 (Posters & Demos)*. Vol. 914. 2012.

[80] John T. Cacioppo, Stephanie Cacioppo, and Richard E. Petty. “The neuroscience of persuasion: A review with an emphasis on issues and opportunities”. In: *Social Neuroscience* (2017), pp. 1–44.

[81] M. W. A. Caminada. “On the issue of reinstatement in argumentation”. In: *Proc. of the 10th European Conference on Logics in Artificial Intelligence (JELIA 2006)*. 2006, pp. 111–123.

[82] M. W. A. Caminada and Gabriella Pigozzi. “On Judgment Aggregation in Abstract Argumentation”. In: *Journal of Autonomous Agents and Multi-Agent Systems* 22.1 (2011), pp. 64–102.

[83] Martin Caminada. “An Algorithm for Computing Semi-stable Semantics”. In: *ECSQARU*. Ed. by Khaled Mellouli. Vol. 4724. LNCS. Springer, 2007, pp. 222–234. ISBN: 978-3-540-75255-4.

[84] Martin Caminada. “On the Issue of Reinstatement in Argumentation”. In: *JELIA*. Ed. by Michael Fisher et al. Vol. 4160. LNCS. Springer, 2006, pp. 111–123. ISBN: 3-540-39625-X.

[85] G. Carenini and J. D. Moore. “Generating and evaluating evaluative arguments”. In: *Artificial Intelligence* 170.11 (2006), pp. 925–952.

[86] Jean Carletta. “Assessing agreement on classification tasks: the kappa statistic”. In: *Computational Linguistics* 22.2 (1996), pp. 249–254. ISSN: 0891-2017. URL: http://dl.acm.org/citation.cfm?id=230386.230390.

[87] Valeria Carofiglio and F. de Rosis. “Combining logical with emotional reasoning in natural argumentation”. In: *9th International Conference on User Modeling. Workshop Proceedings*. Ed. by Conati C, Hudlicka E, and editors Lisetti C. 2003, pp. 9–15.

[88] C. Castelfranchi. “Representation and integration of multiple knowledge sources: issues and questions”. In: *Human & Machine Perception: Information Fusion*. Ed. by V. Cantoni et al. Plenum Press, 1997.

[89] Cristiano Castelfranchi. “Information Agents: The Social Nature of Information and the Role of Trust”. In: *Cooperative Information Agents V, 5th International Workshop, CIA 2001, Modena, Italy, September 6-8, 2001, Proceedings*. Ed. by Matthias Klusch and Franco Zambonelli. Vol. 2182. Lecture Notes in Computer Science. Springer, 2001, pp. 208–210. ISBN: 3-540-42545-4. DOI: 10.1007/3-540-44799-7_22. URL: https://doi.org/10.1007/3-540-44799-7_22.

[90] Cristiano Castelfranchi and Rino Falcone. *Trust Theory: A Socio-Cognitive and Computational Model*. Wiley, 2010.

[91] A. G. Castro et al. “Cognitive support for an argumentative structure during the ontology development process”. In: *Intl. Protege Conference*. 2006.

[92] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. "Bipolarity in Argumentation Graphs: Towards a Better Understanding". In: *Procs of SUM 2011*. Vol. 6929. LNCS. Springer, 2011, pp. 137–148. ISBN: 978-3-642-23962-5.

[93] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. "Coalitions of arguments: A tool for handling bipolar argumentation frameworks". In: *Int. J. Intell. Syst.* 25.1 (2010), pp. 83–109.

[94] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. "On the Acceptability of Arguments in Bipolar Argumentation Frameworks". In: *Proc. of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU), LNCS 3571*. 2005, pp. 378–389.

[95] Claudette Cayrol, Florence Dupin de Saint-Cyr, and Marie-Christine Lagasquie-Schiex. "Change in Abstract Argumentation Frameworks: Adding an Argument". In: *J. Artif. Intell. Res. (JAIR)* 38 (2010), pp. 49–84.

[96] Federico Cerutti, Nava Tintarev, and Nir Oren. "Formal Arguments, Preferences, and Natural Language Interfaces to Humans: an Empirical Evaluation". In: *ECAI 2014 - 21st European Conference on Artificial Intelligence*. 2014, pp. 207–212.

[97] Federico Cerutti et al. "An SCC Recursive Meta-Algorithm for Computing Preferred Labellings in Abstract Argumentation". In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014*. 2014.

[98] F. Cerutti et al. "A SCC Recursive Meta-Algorithm for Computing Preferred Labellings in Abstract Argumentation". In: *Proc. of the 14th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2014)*. 2014, to appear.

[99] Guillaume Chanel et al. "Emotion assessment from physiological signals for adaptation of game difficulty". In: *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 41.6 (2011), pp. 1052–1063.

[100] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: a library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011), p. 27.

[101] Jonathan Chang et al. "Reading tea leaves: How humans interpret topic models". In: *Advances in neural information processing systems*. 2009, pp. 288–296.

[102] Maher Chaouachi, Imène Jraidi, and Claude Frasson. "MENTOR: A Physiologically Controlled Tutoring System". In: *User Modeling, Adaptation and Personalization*. Springer, 2015, pp. 56–67.

[103] Maher Chaouachi, Imène Jraidi, and Claude Frasson. "Modeling mental workload using EEG features for intelligent systems". In: *User modeling, adaption and personalization*. Springer, 2011, pp. 50–61.

[104] C. I. Chesñevar and A.G. Maguitman. "An Argumentative Approach to Assessing Natural Language Usage based on the Web Corpus". In: *European Conference on Artificial Intelligence (ECAI)*. 2004, pp. 581–585.

[105] Carlos Iván Chesñevar et al. "Towards an argument interchange format". In: *Knowledge Eng. Review* 21.4 (2006), pp. 293–316. DOI: 10.1017/S0269888906001044. URL: https://doi.org/10.1017/S0269888906001044.

[106] G. Chierchia and S. McConnell-Ginet. *Meaning and Grammar: An Introduction to Semantics 2nd ed.* Cambridge, MA: MIT Press, 2000.

[107] Philipp Cimiano et al., eds. *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*. Vol. 7882. Lecture Notes in Computer Science. Springer, 2013. ISBN: 978-3-642-38287-1. DOI: 10.1007/978-3-642-38288-8. URL: https://doi.org/10.1007/978-3-642-38288-8.

[108] Helder Coelho, Antônio Carlos da Rocha Costa, and Paulo Trigo. "Moral Minds as Multiple-Layer Organizations". In: *Advances in Artificial Intelligence - IBERAMIA 2010, 12th Ibero-American Conference on AI, Bahıa Blanca, Argentina, November 1-5, 2010. Proceedings*. Ed. by Ángel Fernando Kuri Morales and Guillermo Ricardo Simari. Vol. 6433. Lecture Notes in Computer Science. Springer, 2010, pp. 254–263. ISBN: 978-3-642-16951-9. DOI: 10.1007/978-3-642-16952-6_26. URL: https://doi.org/10.1007/978-3-642-16952-6_26.

[109] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. "Ethical Judgment of Agents' Behaviors in Multi-Agent Systems". In: *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*. Ed. by Catholijn M. Jonker et al. ACM, 2016, pp. 1106–1114. ISBN: 978-1-4503-4239-1. URL: http://dl.acm.org/citation.cfm?id=2937086.

[110] Vincent Conitzer et al. "Moral Decision Making Frameworks for Artificial Intelligence". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. Ed. by Satinder P. Singh and Shaul Markovitch. AAAI Press, 2017, pp. 4831–4835. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14651.

[111] Célia da Costa Pereira and Andrea Tettamanzi. "An Integrated Possibilistic Framework for Goal Generation in Cognitive Agents". In: *AAMAS*. 2010, pp. 1239–1246.

[112] Célia da Costa Pereira, Andrea Tettamanzi, and Serena Villata. "Changing One's Mind: Erase or Rewind?" In: *Procs of IJCAI 2011*. IJCAI/AAAI, 2011, pp. 164–171.

[113] Célia da Costa Pereira, Andrea Tettamanzi, and Serena Villata. "Changing Ones Mind: Erase or Rewind?" In: *Procs of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-2011)*. 2011, pp. 164–171.

[114] Célia da Costa Pereira et al. "Fuzzy Labeling for Abstract Argumentation: An Empirical Evaluation". In: *Scalable Uncertainty Management - 10th International Conference, SUM 2016, Nice, France, September 21-23, 2016, Proceedings*. Ed. by Steven Schockaert and Pierre Senellart. Vol. 9858. Lecture Notes in Computer Science. Springer, 2016, pp. 126–139. ISBN: 978-3-319-45855-7. DOI: 10.1007/978-3-319-45856-4_9. URL: https://doi.org/10.1007/978-3-319-45856-4_9.

[115] S. Coste-Marquis, C. Devred, and P. Marquis. "Prudent Semantics for Argumentation Frameworks". In: *Proc. of the 17th IEEE International Conf. on Tools with Artificial Intelligence (ICTAI 2005)*. IEEE Computer Society. Hong Kong, China, 2005, pp. 568–572.

[116] Sylvie Coste-Marquis et al. "On the merging of Dung's argumentation systems". In: *Artif. Intell.* 171.10-15 (2007), pp. 730–753.

[117] D. Dubois and H. Prade. "An overview of the asymmetric bipolar representation of positive and negative information in possibility theory". In: *Fuzzy Sets Syst.* 160.10 (2009), pp. 1355–1366. ISSN: 0165-0114.

[118] I. Dagan et al. “Recognizing textual entailment: Rational, evaluation and approaches”. In: *JNLE* 15.04 (2009), pp. i–xvii.

[119] Santiago Emanuel Fulladoza Dalibón, Diego César Martinez, and Guillermo Ricardo Simari. “Emotion-directed Argument Awareness for Autonomous Agent Reasoning”. In: *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial* 15.50 (2012), pp. 30–45. URL: http://polar.lsi.uned.es/revista/index.php/ia/article/view/1000.

[120] Clayton Allen Davis et al. “BotOrNot: A System to Evaluate Social Bots”. In: *Proceedings of the 25th International Conference Companion on World Wide Web*. WWW ’16 Companion. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 273–274. ISBN: 978-1-4503-4144-8. DOI: 10.1145/2872518.2889302. URL: https://doi.org/10.1145/2872518.2889302.

[121] J. Daxenberger et al. “What is the Essence of a Claim? Cross-Domain Claim Identification”. In: *EMNLP*. 2017.

[122] B. De Baets, E. Tsiporkova, and R. Mesiar. “Conditioning in possibility theory with strict order norms”. In: *Fuzzy Sets Syst.* 106.2 (1999), pp. 221–229.

[123] James P. Delgrande, Didier Dubois, and Jérôme Lang. “Iterated Revision as Prioritized Merging”. In: *KR*. 2006, pp. 210–220.

[124] David DeSteno et al. “Discrete Emotions and Persuasion: The Role of Emotion-induced Expectancies”. In: *Journal of Personality and Social Psychology* 86 (1 2004), pp. 43–56.

[125] L. R. Dice. “Measures of the amount of ecologic association between species”. In: *Journal of Ecology* 26 (1945), pp. 297–302.

[126] Virginia Dignum. “Responsible Autonomy”. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. Ed. by Carles Sierra. ijcai.org, 2017, pp. 4698–4704. ISBN: 978-0-9992411-0-3. DOI: 10.24963/ijcai.2017/655. URL: https://doi.org/10.24963/ijcai.2017/655.

[127] Jürgen Dix et al. “Research challenges for argumentation”. In: *Computer Science - R&D* 23.1 (2009), pp. 27–34.

[128] Paul Doran et al. “Efficient argumentation over ontology correspondences”. In: *Procs of AAMAS 2009*. 2009, pp. 1241–1242.

[129] Aldo Franco Dragoni and Paolo Giorgini. “Belief Revision Through the Belief-Function Formalism in a Multi-Agent Environment”. In: *ATAL*. Ed. by Jörg P. Müller, Michael Wooldridge, and Nicholas R. Jennings. Vol. 1193. LNCS. Springer, 1996, pp. 103–115. ISBN: 3-540-62507-0.

[130] Didier Dubois and Henri Prade. “A synthetic view of belief revision with uncertain inputs in the framework of possibility theory”. In: *Int. J. Approx. Reasoning* 17.2-3 (1997), pp. 295–324.

[131] P. M. Dung. “On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games”. In: *Artificial Intelligence* 77 (1995), pp. 321–357.

[132] Phan M. Dung. “On the Acceptability of Arguments and its Fundamental Role in Non-monotonic Reasoning, Logic Programming and n-Person Games”. In: *Artif. Intell.* 77.2 (1995), pp. 321–358.

[133] Phan Minh Dung. “On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games”. In: *Artif. Intell.* 77.2 (1995), pp. 321–358.

[134] P. E. Dunne et al. "Weighted argument systems: Basic definitions, algorithms, and complexity results". In: *Artificial Intelligence* 175.2 (2011), pp. 457–486.

[135] E. Durmus and C. Cardie. "Exploting the Role of Prior BEliefs for Argument Persuasion". In: *NAACL*. 2018.

[136] M. Dusmanu, E. Cabrio, and S. Villata. "Argument Mining on Twitter: Arguments, Facts and Sources". In: *EMNLP*. 2017.

[137] Rory Duthie, Katarzyna Budzynska, and Chris Reed. "Mining Ethos in Political Debate". In: *Proceedings of COMMA 2016*. 2016, pp. 299–310. DOI: 10.3233/978-1-61499-686-6-299. URL: http://dx.doi.org/10.3233/978-1-61499-686-6-299.

[138] Camille Dutrey et al. "Local modiÞcations and paraphrases in Wikipedia's revision history". In: *SEPLN Journal* 46 (2011), pp. 51–58.

[139] Nguyen-tuong Duy, R. Peters Jan, and Seeger Matthias. "Local Gaussian Process Regression for Real Time Online Model Learning". In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller et al. Curran Associates, Inc., 2009, pp. 1193–1200. URL: http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2008_0236.pdf.

[140] W. Dvorák and S. Woltran. "On the Intertranslatability of Argumentation Semantics". In: *J. Artificial Intelligence Research (JAIR)* 41 (2011), pp. 445–475.

[141] Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. "On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 2015, pp. 2236–2242.

[142] Douglas C. Engelbart and William K. English. "A Research Center for Augmenting Human Intellect". In: *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*. AFIPS '68 (Fall, part I). San Francisco, California: ACM, 1968, pp. 395–410. DOI: 10.1145/1476589.1476645. URL: http://doi.acm.org/10.1145/1476589.1476645.

[143] Marcelo A. Falappa, Gabriele Kern-Isberner, and Guillermo Simari. "Belief Revision and Argumentation Theory". In: *Argumentation in Artificial Intelligence, I. Rahwan and G. Simari (eds), Springer*. 2009, pp. 341–360.

[144] E. B. Falk, C. N. Cascio, and J. C. Coronel. "Neural prediction of communication-relevant outcomes". In: *Communication Methods and Measures* 9 (1-2 2015), pp. 30–54.

[145] Manaal Faruqui et al. "Retrofitting Word Vectors to Semantic Lexicons". In: *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. Ed. by Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar. The Association for Computational Linguistics, 2015, pp. 1606–1615. ISBN: 978-1-941643-49-5. URL: http://aclweb.org/anthology/N/N15/N15-1184.pdf.

[146] Frederick G. Freeman et al. "Evaluation of a psychophysiologically controlled adaptive automation system, using performance on a tracking task". In: *Applied Psychophysiology and Biofeedback* 25.2 (2000), pp. 103–115.

[147] Frederick Freeman et al. "Evaluation of an adaptive automation system using three EEG indices with a visual tracking task". In: *Biological psychology* 50.1 (1999), pp. 61–76.

[148] N.H. Frijda. *The Emotions*. Studies in Emotion and Social Interaction. Cambridge University Press, 1986. ISBN: 9780521316002. URL: https://books.google.fr/books?id=QkNuuVf-pBMC.

[149] D. M. Gabbay. "Fibring Argumentation Frames". In: *Studia Logica* 93.2-3 (2009), pp. 231–295.

[150] D. M. Gabbay. "Semantics for Higher Level Attacks in Extended Argumentation Frames Part 1: Overview". In: *Studia Logica* 93.2-3 (2009), pp. 357–381.

[151] Dov M. Gabbay, Gabriella Pigozzi, and John Woods. "Controlled Revision - An algorithmic approach for belief revision". In: *J. Log. Comput.* 13.1 (2003), pp. 3–22.

[152] S. A. Gaggl and W. Dvorák. "Stage semantics and the SCC-recursive schema for argumentation semantics". In: *Journal of Logic and Computation* in press (2014).

[153] Diego Gambetta. "Can we trust them?" In: *Trust: Making and breaking cooperative relations* (1990), pp. 213–238.

[154] M.A. Gilbert. "Getting Good Value. Facts, Values, and Goals in Computational Linguistics". In: *Proc. of the International Conference on Computational Science (ICCS), LNCS 2073*. 2001, pp. 989–998.

[155] Michael A. Gilbert. "Emotional Argumentation, or, Why Do Argumentation Theorists Argue with their Mates?" In: *Proceedings of the Third ISSA Conference on Argumentation*. Ed. by F.H. van Eemeren et al. Vol. II. 1995.

[156] Thomas F. Gordon, Henry Prakken, and Douglas Walton. "The Carneades model of argument and burden of proof". In: *Artif. Intell.* 171.10-15 (2007), pp. 875–896.

[157] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37 (3 1969), pp. 424–438.

[158] Nancy Green et al., eds. *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, 2014. URL: http://www.aclweb.org/anthology/W/W14/W14-21.

[159] Samuel W. Greenhouse and Seymour Geisser. "On methods in the analysis of profile data". In: *Psychometrika* 24.2 (1959), pp. 95–112. ISSN: 1860-0980. DOI: 10.1007/BF02289823. URL: http://dx.doi.org/10.1007/BF02289823.

[160] James J. Gross. "Emotion regulation: Affective, cognitive, and social consequences". In: *Psychophysiology* 39 (2002), pp. 281–291.

[161] Kathrin Grosse et al. "Integrating argumentation and sentiment analysis for mining opinions from Twitter". In: *AI Commun.* 28.3 (2015), pp. 387–401. DOI: 10.3233/AIC-140627. URL: http://dx.doi.org/10.3233/AIC-140627.

[162] I. Habernal and I. Gurevych. "Argumentation Mining in User-generated Web Discourse". In: *Comput. Linguist.* 43.1 (2017), pp. 125–179. ISSN: 0891-2017. DOI: 10.1162/COLI_a_00276. URL: https://doi.org/10.1162/COLI_a_00276.

[163] I. Habernal and I. Gurevych. "Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM". In: *ACL*. 2016.

[164] Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. "Argumentation Mining on the Web from Information Seeking Perspective". In: *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlı-Cesena, Italy, July 21-25, 2014*. Ed. by Elena Cabrio, Serena Villata, and Adam Wyner. Vol. 1341. CEUR Workshop Proceedings. CEUR-WS.org, 2014. URL: http://ceur-ws.org/Vol-1341/paper4.pdf.

[165] Ivan Habernal and Iryna Gurevych. "Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse." In: *Proceedings of EMNLP*. 2015, pp. 2127–2137. ISBN: 978-1-941643-32-7. URL: http://dblp.uni-trier.de/db/conf/emnlp/emnlp2015.html#HabernalG15.

[166] S. G. Hart and L. E. Stavenland. "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research". In: *Human Mental Workload*. Ed. by P. A. Hancock and N. Meshkati. Elsevier, 1988. Chap. 7, pp. 139–183. URL: http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000004342%5C_1999205624.pdf.

[167] Stella Heras et al. "How Argumentation can Enhance Dialogues in Social Networks". In: *Computational Model of Arguments (COMMA)*. 2010, pp. 267–274.

[168] Aurélie Herbelot and Eva Maria Vecchi. "Building a shared world: mapping distributional to model-theoretic semantic spaces". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 2015, pp. 22–32.

[169] Michiel Hermans and Benjamin Schrauwen. "Training and Analysing Deep Recurrent Neural Networks". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges et al. 2013, pp. 190–198. URL: http://papers.nips.cc/paper/5166-training-and-analysing-deep-recurrent-neural-networks.

[170] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[171] Mohammed E. Hoque, Daniel McDuff, and Rosalind W. Picard. "Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles." In: *T. Affective Computing* 3.3 (2012), pp. 323–334. URL: http://dblp.uni-trier.de/db/journals/taffco/taffco3.html#HoqueMP12.

[172] Alexander C. Huk and Miriam L. R. Meister. "Neural correlates and neural computations in posterior parietal cortex during perceptual decision-making". In: *Front. Integr. Neurosci.* 6 (86 2012).

[173] Anthony Hunter. "Computational Persuasion with Applications in Behaviour Change". In: *Proc. of COMMA 2016*. 2016, pp. 5–18. DOI: 10.3233/978-1-61499-686-6-5. URL: http://dx.doi.org/10.3233/978-1-61499-686-6-5.

[174] Anthony Hunter and Matthew Williams. "Aggregating evidence about the positive and negative effects of treatments". In: *Artificial Intelligence in Medicine* 56.3 (2012), pp. 173–190. DOI: 10.1016/j.artmed.2012.09.004. URL: http://dx.doi.org/10.1016/j.artmed.2012.09.004.

[175] Carroll E Izard. *The psychology of emotions*. Springer Science & Business Media, 1991.

[176] Hadassa Jakobovits and Dirk Vermeir. "Robust Semantics for Argumentation Frameworks". In: *J. Log. Comput.* 9.2 (1999), pp. 215–261.

[177] T. Janhunen et al. "Modularity Aspects of Disjunctive Stable Models". In: *J. of Artificial Intelligence Research (JAIR)* 35 (2009), pp. 813–857.

[178] Mathilde Janier, John Lawrence, and Chris Reed. "OVA+: an Argument Analysis Interface". In: *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*. Ed. by Simon Parsons et al. Vol. 266. Frontiers in Artificial Intelligence and Applications. IOS Press, 2014, pp. 463–464. ISBN: 978-1-61499-435-0. DOI: 10.3233/978-1-61499-436-7-463. URL: https://doi.org/10.3233/978-1-61499-436-7-463.

[179] J. Janssen, M. De Cock, and D. Vermeir. "Fuzzy Argumentation Frameworks". In: *Procedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2008)*. 2008, pp. 513–520.

[180] Oliver P. John and Sanjay Srivastava. "The Big-Five trait taxonomy: History, measurement, and theoretical perspectives". In: *Handbook of personality: Theory and Research*. Guilford Press, 1999, pp. 102–138.

[181] Samuel G. B. Johnson, Thomas Merchant, and Frank Keil. "Argument Scope in Inductive Reasoning: Evidence for an Abductive Account of Induction". In: *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015*. Ed. by David C. Noelle et al. cognitivesciencesociety.org, 2015. ISBN: 978-0-9911967-2-2. URL: https://mindmodeling.org/cogsci2015/papers/0181/index.html.

[182] Imène Jraidi, Maher Chaouachi, and Claude Frasson. "A dynamic multimodal approach for assessing learner's interaction experience". In: *Proceedings of the 15th ACM on International Conference on multimodal interaction*. ACM, 2013, pp. 271–278.

[183] Imene Jraidi and Claude Frasson. "Student's Uncertainty Modeling through a Multimodal Sensor-Based Approach". In: *Educational Technology & Society* 16.1 (2013), pp. 219–230. URL: http://www.ifets.info/download_pdf.php?j_id=58&a_id=1329.

[184] Trabulsi Julia, Manuel Garcia-Garcia, and Michael E. Smith. "Consumer neuroscience: A method for optimising marketing communication". In: *Journal of Cultural Marketing Strategy* 1 (2015), pp. 80–89.

[185] I. Kant and M.J. Gregor. *Practical Philosophy*. Kant, Immanuel, 1724-1804. Works. Engl. 1992. Cambridge University Press, 1999. ISBN: 9780521654081. URL: https://books.google.it/books?id=0hCsbUjFiBwC.

[186] Andrej Karpathy, Justin Johnson, and Fei-Fei Li. "Visualizing and Understanding Recurrent Networks". In: *CoRR* abs/1506.02078 (2015). URL: http://arxiv.org/abs/1506.02078.

[187] K. Kellermann et al. "The conversation MOP: Scenes in the stream of discourse". In: *Discourse Processes* 12 (1 1989), pp. 27–61.

[188] Diederik Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[189] M. Kouylekov and M. Negri. "An Open-Source Package for Recognizing Textual Entailment". In: *ACL System Demonstrations*. 2010, pp. 42–47.

[190] Loredana Laera et al. “Argumentation over ontology correspondences in MAS”. In: *Procs of AAMAS 2007*. 2007, pp. 1–8.

[191] Thomas K. Landauer et al. “How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans.” In: *Proc. of CSS*. 1997, pp. 412–417.

[192] J. Richard Landis and Gary G. Koch. “The Measurement of Observer Agreement for Categorical Data”. In: *Biometrics* 33.1 (1977), pp. 159–174. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2529310.

[193] C. Lange et al. “Expressing Argumentative Discussions in Social Media Sites”. In: *SDoW*. 2008.

[194] R.S. Lazarus. *Emotion and Adaptation*. Oxford University Press, 1994. ISBN: 9780195092660. URL: https://books.google.fr/books?id=tTdIlwpxtWsC.

[195] Y. Y. Lee and S. Hsieh. “Classifying different emotional states by means of EEG-based functional connectivity patterns”. In: *PloS one* 9.4 (2014).

[196] Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. “Rationalizing Neural Predictions”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. Ed. by Jian Su, Xavier Carreras, and Kevin Duh. The Association for Computational Linguistics, 2016, pp. 107–117. ISBN: 978-1-945626-25-8. URL: http://aclweb.org/anthology/D/D16/D16-1011.pdf.

[197] J. Leite and J. Martins. “Social Abstract Argumentation”. In: *Proc. of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*. 2011, pp. 2287–2292.

[198] Ran Levy et al. “Context Dependent Claim Detection”. In: *Proceedings of COLING*. 2014, pp. 1489–1500. URL: http://aclweb.org/anthology/C/C14/C14-1141.pdf.

[199] Jiwei Li et al. “Visualizing and Understanding Neural Models in NLP”. In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. The Association for Computational Linguistics, 2016, pp. 681–691. ISBN: 978-1-941643-91-4. URL: http://aclweb.org/anthology/N/N16/N16-1082.pdf.

[200] B. Liao and H. Huang. “Partial semantics of argumentation: basic properties and empirical results”. In: *J. of Logic and Computation* 23.3 (2013), pp. 541–562.

[201] B. Liao, L. Jin, and R. C. Koons. “Dynamics of argumentation systems: A division-based method”. In: *Artificial Intelligence* 175 (2011), pp. 1790–1814.

[202] Churn-Jung Liau. “Belief, information acquisition, and trust in multi-agent systems–A modal logic formulation”. In: *Artif. Intell.* 149.1 (2003), pp. 31–60.

[203] V. Lifschitz, D. Pearce, and A. Valverde. “Strongly Equivalent Logic Programs”. In: *ACM Trans. on Computational Logic* 2.4 (2001), pp. 526–541.

[204] M. Lippi and P. Torroni. “MARGOT: A web server for argumentation mining”. In: *Expert Systems with Applications* 65 (2016), pp. 292–303.

[205] Marco Lippi and Paolo Torroni. “Argument Mining from Speech: Detecting Claims in Political Debates”. In: *Proceedings AAAI*. 2016, pp. 2979–2985. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12164.

[206] Marco Lippi and Paolo Torroni. "Argumentation Mining: State of the Art and Emerging Trends". In: *ACM Trans. Internet Techn.* 16.2 (2016), p. 10. DOI: 10.1145/2850417. URL: http://doi.acm.org/10.1145/2850417.

[207] Marco Lippi and Paolo Torroni. "Context-Independent Claim Detection for Argument Mining". In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. Ed. by Qiang Yang and Michael Wooldridge. AAAI Press, 2015, pp. 185–191. ISBN: 978-1-57735-738-4. URL: http://ijcai.org/papers15/Abstracts/IJCAI15-033.html.

[208] B. Liu. *Sentiment Analysis and Opinion Mining*. Synthesis digital library of engineering and computer science. Morgan & Claypool, 2012. ISBN: 9781608458844. URL: https://books.google.fr/books?id=Gt8g72e6MuEC.

[209] Martyn Lloyd-Kelly and Adam Wyner. "Arguing about Emotion". In: *Advances in User Modeling - UMAP 2011 Workshops*. 2011, pp. 355–367.

[210] Vanessa Lopez et al. "Is Question Answering fit for the Semantic Web?: A survey". In: *Semantic Web* 2.2 (2011), pp. 125–155.

[211] Emiliano Lorini. "On the Logical Foundations of Moral Agency". In: *Deontic Logic in Computer Science - 11th International Conference, DEON 2012, Bergen, Norway, July 16-18, 2012. Proceedings*. 2012, pp. 108–122.

[212] Emiliano Lorini and Robert Demolombe. "From Binary Trust to Graded Trust in Information Sources: A Logical Perspective". In: *Trust in Agent Societies, 11th International Workshop (TRUST-2008)*. Vol. 5396. Lecture Notes in Computer Science. Springer, 2008, pp. 205–225.

[213] Bernardo Magnini et al. "The Excitement Open Platform for Textual Inferences". In: *Proceedings of ACL (System Demonstrations)*. 2014, pp. 43–48.

[214] Christopher D. Manning et al. "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of ACL (System Demonstrations)*. 2014, pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010.

[215] M.C. De Marneffe, A.N. Rafferty, and C.D. Manning. "Finding contradictions in text". In: *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2008.

[216] Paul-Amaury Matt, Maxime Morge, and Francesca Toni. "Combining statistics and arguments to compute trust". In: *Procs of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2010)*. 2010, pp. 209–216.

[217] Paul-Amaury Matt and Francesca Toni. "A Game-Theoretic Measure of Argument Strength for Abstract Argumentation". In: *Procs of JELIA 2008*. Vol. 5293. LNCS. Springer, 2008, pp. 285–297.

[218] J. Mauchly. "Significance Test for Sphericity of a Normal n-Variate Distribution". In: *Ann. Math. Stat.* 11.2 (1940), pp. 204–209. DOI: 10.1214/aoms/1177731915. URL: http://dx.doi.org/10.1214/aoms/1177731915.

[219] Aurelien Max and Guillaume Wisniewski. "Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History". In: *LREC*. 2010.

[220] Pablo N. Mendes, Hannes Muhleisen, and Christian Bizer. "Sieve: linked data quality assessment and fusion". In: *Procs of the Joint EDBT/ICDT Workshops*. ACM, 2012, pp. 116–123.

[221] Pablo Mendes, Max Jakob, and Christian Bizer. "DBpedia: A Multilingual Cross-domain Knowledge Base". In: *Procs of LREC 2012*. ELRA, 2012.

[222] Stefano Menini and Sara Tonelli. "Agreement and Disagreement: Comparison of Points of View in the Political Domain." In: *Proceedings of COLING*. 2016.

[223] Stefano Menini et al. "Never Retreat, Never Retract: Argumentation Analysis for Political Speeches". In: *AAAI-2018*. 2018.

[224] Stefano Menini et al. "Topic-based agreement and disagreement in US electoral manifestos". English. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, 2017, pp. 2928–2934. URL: http://ub-madoc.bib.uni-mannheim.de/42490/.

[225] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[226] Brent Daniel Mittelstadt and Luciano Floridi. "Transparent, explainable, and accountable AI for robotics". In: *Science Robotics* 2.6 (2017). DOI: 10.1126/scirobotics.aan6080. URL: https://doi.org/10.1126/scirobotics.aan6080.

[227] Raquel Mochales and Marie-Francine Moens. "Argumentation mining". In: *Artificial Intelligence and Law* 19.1 (2011), pp. 1–22. ISSN: 1572-8382. DOI: 10.1007/s10506-010-9104-x. URL: http://dx.doi.org/10.1007/s10506-010-9104-x.

[228] S. Modgil. "Reasoning about preferences in argumentation frameworks". In: *Artificial Intelligence* 173.9-10 (2009), pp. 901–934.

[229] S. Modgil and T. J. M. Bench-Capon. "Integrating Object and Meta-level Value Based Argumentation". In: *Proc. of the 2nd Int. Conf. on Computational Models of Argument (COMMA 2008)*. 2008, pp. 240–251.

[230] S. Modgil and T.J.M Bench-Capon. *Metalevel argumentation*. Tech. rep. www.csc.liv.ac.uk/research/techreports/techreports.html, 2009.

[231] M.F. Moens et al. "Automatic Detection of Arguments in Legal Texts". In: *Proc. of the International Conference on Artificial Intelligence and Law (ICAIL)* (2007), pp. 225–230.

[232] C. Monz and M. de Rijke. "Light-Weight Entailment Checking for Computational Semantics". In: *Proc. Inference in Computational Semantics (ICoS-3)*. 2001, pp. 59–72.

[233] Nona Naderi and Graeme Hirst. "Argumentation mining in parliamentary discourse". In: *Proceedings of the 15th Workshop on Computational Models of Natural Argument (CMNA-2015)*. 2015.

[234] Fahd Saud Nawwab, Paul E. Dunne, and Trevor J. M. Bench-Capon. "Exploring the Role of Emotions in Rational Decision Making". In: *Computational Models of Argument: Proceedings of COMMA 2010*. 2010, pp. 367–378.

[235] V. Niculae, J. Park, and C. Cardie. "Argument Mining with Structured SVMs and RNNs". In: *ACL*. 2017. DOI: 10.18653/v1/P17-1091. URL: http://www.aclweb.org/anthology/P17-1091.

[236] Farhad Nooralahzadeh et al. “Adapting Semantic Spreading Activation to Entity Linking in Text”. In: *Natural Language Processing and Information Systems - 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings*. 2016, pp. 74–90. URL: http://dx.doi.org/10.1007/978-3-319-41754-7%5C_7.

[237] Farid Nouioua and Vincent Risch. “Argumentation Frameworks with Necessities”. In: *Proc. of the 5th International Conference Scalable Uncertainty Management (SUM), LNCS 6929*. 2011, pp. 163–176.

[238] Farid Nouioua and Vincent Risch. “Bipolar Argumentation Frameworks with Specialized Supports”. In: *Proc. of the 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE Computer Society, 2010, pp. 215–218.

[239] E. Oikarinen and T. Janhunen. “Modular Equivalence for Normal Logic Programs”. In: *Proc. of the 17th European Conf. on Artificial Intelligence (ECAI 2006)*. 2006, pp. 412–416.

[240] E. Oikarinen and S. Woltran. “Characterizing strong equivalence for argumentation frameworks”. In: *Artificial Intelligence* 175 (2011), pp. 1985–2009.

[241] Nir Oren and Timothy J. Norman. “Semantics for Evidence-Based Argumentation”. In: *Proc. of the International Conference on Computational Models of Argument (COMMA), Frontiers in Artificial Intelligence and Applications 172*. 2008, pp. 276–284.

[242] Nir Oren, Chris Reed, and Michael Luck. “Moving Between Argumentation Frameworks”. In: *Proc. of the International Conference on Computational Models of Argument (COMMA), Frontiers in Artificial Intelligence and Applications 216*. 2010, pp. 379–390.

[243] A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, 1988.

[244] Beyond brain mapping: using neural measures to predict real-world outcomes. In: *Curr. Dir. Psychol. Sci.* 22 (2013), pp. 45–50.

[245] P. Baroni, M. Giacomin, and G. Guida. “SCC-recursiveness: a general schema for argumentation semantics”. In: *Artificial Intelligence Journal* 168.1-2 (2005), pp. 165–210.

[246] F. Paglieri and C. Castelfranchi. “Arguing on the Toulmin model”. In: Berlin, Springer, 2006. Chap. The Toulmin Test: Framing argumentation within belief revision theories, pp. 359–377.

[247] Fabio Paglieri et al. “Trusting the messenger because of the message: feedback dynamics from information quality to source evaluation”. In: *Computational & Mathematical Organization Theory* 20.2 (2014), pp. 176–194. DOI: 10.1007/s10588-013-9166-x. URL: https://doi.org/10.1007/s10588-013-9166-x.

[248] Raquel Mochales Palau and Marie-Francine Moens. “Argumentation mining”. In: *Artif. Intell. Law* 19.1 (2011), pp. 1–22. DOI: 10.1007/s10506-010-9104-x. URL: http://dx.doi.org/10.1007/s10506-010-9104-x.

[249] R. Parasuraman and D. Caggiano. “Mental workload”. In: *Encyclopedia of the human brain* 3 (2002), pp. 17–27.

[250] J. Park and C. Cardie. “Identifying appropriate support for propositions in online user comments”. In: *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, 2014.

[251] Joonsuk Park, Cheryl Blake, and Claire Cardie. "Toward machine-assisted participation in eRulemaking: an argumentation model of evaluability". In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law, ICAIL 2015, San Diego, CA, USA, June 8-12, 2015*. 2015, pp. 206–210. DOI: 10.1145/2746090.2746118. URL: http://doi.acm.org/10.1145/2746090.2746118.

[252] Simon Parsons, Peter McBurney, and Elizabeth Sklar. "Reasoning about Trust using Argumentation: A position paper". In: *Procs. of the 7th International Workshop on Argumentation in Multi-Agent Systems (ArgMAS-2010)*. 2010.

[253] Simon Parsons et al. "Argumentation-based reasoning in agents with varying degrees of trust". In: *Procs of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2011)*. 2011, pp. 879–886.

[254] Simon Parsons et al., eds. *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*. Vol. 266. Frontiers in Artificial Intelligence and Applications. IOS Press, 2014. ISBN: 978-1-61499-435-0.

[255] A. Peldszus and M. Stede. "Joint prediction in MST-style discourse parsing for argumentation mining". In: *EMNLP*. 2015.

[256] Andreas Peldszus and Manfred Stede. "From Argument Diagrams to Argumentation Mining in Texts: A Survey". In: *IJCINI* 7.1 (2013), pp. 1–31. DOI: 10.4018/jcini.2013010101. URL: http://dx.doi.org/10.4018/jcini.2013010101.

[257] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[258] I. Persing and V. Ng. "Why Can't You Convince Me? Modeling Weaknesses in Unpersuasive Arguments". In: *IJCAI*. 2017.

[259] M. Pinkal. "Logic and Lexicon: the semantics of the indefinite". In: *Studies in linguistics and philosophy* 56 (1995).

[260] Alan T. Pope, Edward H. Bogart, and Debbie S. Bartolome. "Biocybernetic system evaluates indices of operator engagement in automated task". In: *Biological psychology* 40.1 (1995), pp. 187–195.

[261] Henri Prade. "A Qualitative Bipolar Argumentative View of Trust". In: *Scalable Uncertainty Management, First International Conference (SUM-2007)*. Vol. 4772. Lecture Notes in Computer Science. Springer, 2007, pp. 268–276.

[262] H. Prakken. "Reconstructing Popov v. Hayashi in a framework for argumentation with structured arguments and Dungean semantics". In: *Artificial Intelligence and Law* 20.1 (2012), pp. 57–82.

[263] Henry Prakken. "An abstract framework for argumentation with structured arguments". In: *Argument & Computation* 1.2 (2010), pp. 93–124. DOI: 10.1080/19462160903564592. URL: https://doi.org/10.1080/19462160903564592.

[264] Matthieu Quignard and Michael Baker. "Modelling argumentation and belief revision in agents interactions". In: *European Conference on Cognitive Science*. 1997, pp. 85–90.

[265] A. Calvo Rafael and D'Mello Sidney. "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications". In: *IEEE Transactions on Affective Computing* 1.1 (2010), pp. 18–37. ISSN: 1949-3045. DOI: http://doi.ieeecomputersociety.org/10.1109/T-AFFC.2010.1.

[266] Iyad Rahwan and Guillermo R. Simari. *Argumentation in Artificial Intelligence*. 1st. Springer Publishing Company, Incorporated, 2009. ISBN: 0387981969, 9780387981963.

[267] Iyad Rahwan et al. "Behavioral Experiments for Assessing the Abstract Argumentation Semantics of Reinstatement". In: *Cognitive Science* 34.8 (2010), pp. 1483–1502.

[268] Iyad Rahwan et al. "Representing and classifying arguments on the Semantic Web". In: *Knowledge Eng. Review* 26.4 (2011), pp. 487–511.

[269] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN: 026218253X.

[270] C. Reed and F. Grasso. "Recent advances in computational models of natural argument". In: *Int. J. Intell. Syst.* 22.1 (2007), pp. 1–15.

[271] C. Reed and G. Rowe. "Araucaria: Software for Argument Analysis, Diagramming and Representation". In: *Int. Journal on Artificial Intelligence Tools* 13.4 (2004), pp. 961–980.

[272] T. Rienstra et al. "Multi-sorted argumentation frameworks". In: *Theory and Applications of Formal Argumentation - First Int. Workshop (TAFA 2011). Revised Selected Papers*. Vol. 7132. Lecture Notes in Computer Science. Springer, 2011, pp. 231–245.

[273] Tim Rocktaschel et al. "Reasoning about Entailment with Neural Attention". In: *International Conference on Learning Representations*. 2016.

[274] Ana Rojo. *Step by Step: A Course in Contrastive Linguistics and Translation*. Ed. by Karl A. Bernhardt Graeme Davis. Peter Lang, 2009.

[275] L. Romano et al. "Investigating a Generic Paraphrase-Based Approach for Relation Extraction". In: *Proc. of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 2006, pp. 409–416.

[276] Ariel Rosenfeld and Sarit Kraus. "Strategical Argumentative Agent for Human Persuasion". In: *Proc. of ECAI*. 2016, pp. 320–328. DOI: 10.3233/978-1-61499-672-9-320. URL: http://dx.doi.org/10.3233/978-1-61499-672-9-320.

[277] W.D. Ross and W.R. Roberts. *Rhetoric - Aristotle*. Cosimo Classics Philosophy. 2010. ISBN: 9781616403072. URL: https://books.google.fr/books?id=t8Rl2z8XYywC.

[278] Paul Rozin and Edward B. Royzman. "Negativity Bias, Negativity Dominance, and Contagion". In: *Personality and Social Psychology Review* 5.4 (2001), pp. 296–320. DOI: 10.1207/S15327957PSPR0504_2. eprint: https://doi.org/10.1207/S15327957PSPR0504_2. URL: https://doi.org/10.1207/S15327957PSPR0504_2.

[279] Paul Rozin and Edward B. Royzman. "Negativity bias, negativity dominance, and contagion". In: *Personality and social psychology review* 5.4 (2001), pp. 296–320.

[280] Alexander M. Rush, Sumit Chopra, and Jason Weston. "A Neural Attention Model for Abstractive Sentence Summarization". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 2015, pp. 379–389.

[281] Stuart J. Russell. "Provably Beneficial Artificial Intelligence". In: *Exponential Life, The Next Step* (2017).

[282] Stuart J. Russell et al. "Letter to the Editor: Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter". In: *AI Magazine* 36.4 (2015). URL: http://www.aaai.org/ojs/index.php/aimagazine/article/view/2621.

[283] Cassia Trojahn dos Santos and Jerome Euzenat. "Consistency-driven argumentation for alignment agreement". In: *Procs of OM 2010, CEUR Workshop Proceedings 689*. 2010.

[284] R. Saxe and A. Wexler. "Making sense of another mind: The role of the right temporoparietal junction". In: *Neuropsychologia* 43 (10 2005), pp. 1391–9.

[285] Klaus R. Scherer. "What are emotions? And how can they be measured?" In: *Social science information* 44(4) (2005), pp. 695–729.

[286] J. Schneider, T. Groza, and A. Passant. "A Review of Argumentation for the Social Semantic Web". In: *Semantic Web J.* (2011).

[287] John Searle. "Watson Doesn't Know it Won on "Jeopardy"". In: *Wall Street Journal* (2011).

[288] Richard Socher et al. "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of EMNLP*. Vol. 1631. 2013, p. 1642.

[289] C. Stab and I. Gurevych. "Parsing Argumentation Structures in Persuasive Essays". In: *Comput. Linguist.* 43.3 (2017), pp. 619–659. ISSN: 0891-2017. DOI: 10.1162/COLI_a_00295. URL: https://doi.org/10.1162/COLI_a_00295.

[290] Christian Stab and Iryna Gurevych. "Identifying Argumentative Discourse Structures in Persuasive Essays". In: *Proceedings of EMNLP*. 2014, pp. 46–56. URL: http://aclweb.org/anthology/D/D14/D14-1006.pdf.

[291] Christian Stab and Iryna Gurevych. "Parsing Argumentation Structures in Persuasive Essays". In: *CoRR* abs/1604.07370 (2016). URL: http://arxiv.org/abs/1604.07370.

[292] Oliviero Stock, Marco Guerini, and Fabio Pianesi. "Ethical Dilemmas for Adaptive Persuasion Systems". In: *Proc. of AAAI*. 2016, pp. 4157–4162.

[293] Ruben Stranders, Mathijs de Weerdt, and Cees Witteveen. "Fuzzy Argumentation for Trust". In: *Computational Logic in Multi-Agent Systems, 8th International Workshop (CLIMA VIII)*. Vol. 5056. Lecture Notes in Computer Science. Springer, 2007, pp. 214–230.

[294] Daniel Szafir and Bilge Mutlu. "ARTFul: adaptive review technology for flipped learning". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2013, pp. 1001–1010.

[295] Daniel Szafir and Bilge Mutlu. "Pay attention!: designing adaptive agents that monitor and improve user engagement". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2012, pp. 11–20.

[296] Yuqing Tang et al. "A system of argumentation for reasoning about trust". In: *Procs. of the 8th European Workshop on Multi-Agent Systems (EUMAS-2010)*. 2010.

[297] M. Teplan. "Fundamentals of EEG measurement". In: *Measurement Science Review* 2 (2002).

[298] M. Teruel et al. "Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines". In: *LREC*. 2018.

[299] Simone Teufel, Advaith Siddharthan, and Colin R. Batchelor. "Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics". In: *Proceedings of EMNLP*. 2009, pp. 1493–1502. URL: http://www.aclweb.org/anthology/D09-1155.

[300] Matthias Thimm and Serena Villata. "System Descriptions of the First International Competition on Computational Models of Argumentation (ICCMA'15)". In: *CoRR* abs/1510.05373 (2015). URL: http://arxiv.org/abs/1510.05373.

[301] Matthias Thimm and Serena Villata. "The first international competition on computational models of argumentation: Results and analysis". In: *Artif. Intell.* 252 (2017), pp. 267–294. DOI: 10.1016/j.artint.2017.08.006. URL: https://doi.org/10.1016/j.artint.2017.08.006.

[302] S.E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2003. ISBN: 9781139442305. URL: https://books.google.fr/books?id=53whAwAAQBAJ.

[303] Mauro Vallati, Federico Cerutti, and Massimiliano Giacomin. "Argumentation Frameworks Features: an Initial Study". In: *ECAI 2014 - 21st European Conference on Artificial Intelligence*. 2014, pp. 1117–1118.

[304] B. Verheij. "Argumed - a template-based argument mediation system for lawyers and legal knowledge based systems". In: *Proc. of the 11th International Conference on Legal Knowledge and Information Systems (JURIX)*. 1998, pp. 113–130.

[305] B. Verheij. "Artificial argument assistants for defeasible argumentation". In: *Artif. Intell.* 150.1-2 (2003), pp. 291–324.

[306] B. Verheij. "Two approaches to dialectical argumentation: admissible sets and argumentation stages". In: *Proc. of the Eighth Dutch Conf. on Artificial Intelligence (NAIC'96)*. 1996, pp. 357–368.

[307] Serena Villata. "Meta-Argumentation for Multiagent Systems: Coalition Formation, Merging Views, Subsumption Relation and Dependence Networks". PhD thesis. University of Turin, 2010.

[308] Serena Villata, Guido Boella, and Leendert van der Torre. "Argumentation Patterns". In: *Proceedings of the 8th International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2011)*. 2011, pp. 133–150.

[309] Serena Villata et al. "A socio-cognitive model of trust using argumentation theory". In: *Int. J. Approx. Reasoning* 54.4 (2013), pp. 541–559. DOI: 10.1016/j.ijar.2012.09.001. URL: https://doi.org/10.1016/j.ijar.2012.09.001.

[310] Serena Villata et al. "Arguing about the Trustworthiness of the Information Sources". In: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 11th European Conference (ECSQARU-2011)*. Vol. 6717. Lecture Notes in Computer Science. Springer, 2011, pp. 74–85.

[311] Serena Villata et al. "Assessing Persuasion in Argumentation through Emotions and Mental States". In: *Proceedings of the Thirty-first International Florida Artificial Intelligence Research Society Conference, FLAIRS 2018*. 2018.

[312] Serena Villata et al. "Emotions and personality traits in argumentation: An empirical evaluation". In: *Argument & Computation* 8.1 (2017), pp. 61–87.

[313] P. Vuilleumier. "How brains beware: neural mechanisms of emotional attention". In: *Trends Cogn. Sci.* 9 (12 2005), pp. 585–594.

[314] Henning Wachsmuth et al. “Modeling Review Argumentation for Robust Sentiment Analysis”. In: *Proceedings of COLING*. 2014, pp. 553–564. URL: http://aclweb.org/anthology/C/C14/C14-1053.pdf.

[315] H. Wachsmuth et al. “Argumentation Quality Assessment: Theory vs. Practice”. In: *ACL*. 2017.

[316] Douglas Walton. “Explanations and Arguments Based on Practical Reasoning”. In: *Explanation-aware Computing, Papers from the 2009 IJCAI Workshop, Pasadena, California , USA, July 11-12, 2009*. Ed. by Thomas Roth-Berghofer, Nava Tintarev, and David B. Leake. 2009, pp. 72–83.

[317] Douglas Walton. *The Place of Emotion in Argument*. Pennsylvania State University Press, University Park, 1992.

[318] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[319] Lu Wang and Claire Cardie. “Improving Agreement and Disagreement Identification in Online Discussions with A Socially-Tuned Sentiment Lexicon”. In: *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@ACL 2014, June 27, 2014, Baltimore, Maryland, USA*. 2014, pp. 97–106. URL: http://aclweb.org/anthology/W/W14/W14-2617.pdf.

[320] Weihong Wang et al. “Indexing cognitive workload based on pupillary response under luminance and emotional changes”. In: *18th International Conference on Intelligent User Interfaces, IUI, 13, Santa Monica, CA, USA, March 19-22, 2013*. Ed. by Jihie Kim, Jeffrey Nichols, and Pedro A. Szekely. ACM, 2013, pp. 247–256. ISBN: 978-1-4503-1965-2. DOI: 10.1145/2449396.2449428. URL: http://doi.acm.org/10.1145/2449396.2449428.

[321] Yonghong Wang and Munindar P. Singh. “Formal Trust Model for Multiagent Systems”. In: *Procs of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007)*. 2007, pp. 1551–1556.

[322] Thomas Wehrle and Susanne Kaiser. “Emotion and facial expression”. In: *Affective interactions*. Springer Berlin Heidelberg, 2000, pp. 49–63.

[323] Joseph Jay Williams and Tania Lombrozo. “The Role of Explanation in Discovery and Generalization: Evidence From Category Learning”. In: *Cognitive Science* 34.5 (2010), pp. 776–806. DOI: 10.1111/j.1551-6709.2010.01113.x. URL: http://dx.doi.org/10.1111/j.1551-6709.2010.01113.x.

[324] A. Z. Wyner and T. J. M. Bench-Capon. “Argument Schemes for Legal Case-based Reasoning”. In: *Proc. of the 20th Annual Conf. on Legal Knowledge and Information Systems (JURIX 2007)*. 2007, pp. 139–149.

[325] A. Wyner and T. van Engers. “A framework for enriched, controlled on-line discussion forums for e-government policy-making”. In: *eGov*. 2010.

[326] Elif Yamangil and Rani Nelken. “Mining Wikipedia Revision Histories for Improving Sentence Compression”. In: *ACL (Short Papers)*. 2008, pp. 137–140.

[327] Qiang Yang and Michael Wooldridge, eds. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. AAAI Press, 2015. ISBN: 978-1-57735-738-4.

[328] Mark Yatskar et al. “For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia”. In: *HLT-NAACL*. 2010, pp. 365–368.

[329] L. A. Zadeh. “Fuzzy Sets as a Basis for a Theory of Possibility”. In: *Fuzzy Sets and Systems* 1 (1978), pp. 3–28.

[330] F.M. Zanzotto and M. Pennacchiotti. “Expanding textual entailment corpora from Wikipedia using co-training”. In: *The People’s Web Meets NLP Workshop*. 2010.