Argument Mining on Clinical Trials

Tobias Mayer

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Fouille d'arguments à partir des essais cliniques

Tobias MAYER

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis

Présentée en vue de l'obtention du grade de docteur en Informatique **d'**Université Côte d'Azur
Dirigée par : Serena VILLATA *et* Céline POUDAT
Co-encadrée par : Elena CABRIO
Soutenue le : 17.12.2020

Devant le jury, composé de :
Président du jury : Michel RIVEILL,
Full Professor, Université Côte d'Azur
Rapporteurs:
Iryna GUREVYCH,
Full Professor, TU Darmstadt
Smaranda MURESAN,
Associate Professor, Columbia University
Examinateur: Anthony HUNTER,
Full Professor, University College London

Inria

i3S sophia antipolis

Fouille d'arguments à partir des essais cliniques

Argument Mining on Clinical Trials

COMPOSITION DU JURY

Président du jury:
Michel RIVEILL, Professeur, Université Côte d'Azur

Rapporteurs:
Iryna GUREVYCH, Professeur, TU Darmstadt
Smaranda MURESAN, Associate Professor, Columbia University

Examinateur:
Anthony HUNTER, Professeur, University College London

Invité:
Elena CABRIO, Maître de Conférences, Université Côte d'Azur

Directeurs de thèse:
Serena VILLATA, Charge de recherche, HDR, Université Côte d'Azur
Céline POUDAT, Maître de Conférences, Université Côte d'Azur

Résumé

Ces dernières années, le domaine de la e-santé a vu un intérêt croissant pour la définition de systèmes intelligents ayant le but d'accompagner les cliniciens dans leurs tâches et leurs activités quotidiennes. D'ailleurs, cela inclut de nouveaux systèmes pour le domaine de la médecine basée sur les preuves. Ce dernier repose sur le principe de l'évaluation critique des preuves médicales et de la combinaison de ces preuves de haute qualité avec l'expérience clinique individuelle du praticien par rapport à la situation d'un patient pour obtenir le meilleur résultat possible. La plupart des systèmes intelligents proposés visent soit à extraire des informations sur la qualité des preuves issues des essais cliniques, de directives cliniques ou des dossiers de santé électroniques, soit à aider dans les processus de prise de décision, sur la base de cadres de raisonnement. Le travail de cette thèse va au-delà de l'état de l'art des systèmes d'extraction d'informations actuellement proposés dans ce contexte. Il utilise des méthodes d'analyse d'arguments pour extraire et classifier les composants d'argumentation (c'est-à-dire les preuves et les conclusions d'un essai clinique) et leurs relations (c'est-à-dire le support et l'attaque). Un cadre de fouille d'arguments (Argument Mining) est proposé et amélioré pour intégrer des informations supplémentaires inspirées par les cadres biomédicaux courants pour l'analyse des essais cliniques. Ces extensions comprennent la détection des éléments PICO et un module d'analyse des résultats pour identifier et classer les effets (c'est-à-dire améliorés, augmentés, diminués, pas de différence, pas d'occurrence) d'une intervention sur le résultat de l'essai. Dans ce contexte, un jeu de données, composé de 660 résumés d'essais cliniques dans la base de données MEDLINE, a été annoté, en résultant dans le construction d'un jeu de données étiquetées qui inclut 4198 composants d'argumentation, 2601 relations d'argumentation et 3351 résultats d'intervention sur cinq maladies différentes (néoplasme, glaucome, hépatite, diabète, hypertension). Diverses approches d'apprentissage automatique et profond allant des SVM aux architectures récentes basées sur les réseaux de neurones ont été expérimentées, obtenant un F1 macro de 0,87 pour la détection de composants d'argumentation et de 0,68 pour la prédiction des relation d'argumentation, surpassant les résultats obtenus pas les systèmes de detection d'arguments dans l'état de l'art. De plus, une demo d'un système, appelé ACTA, a été développée pour démontrer l'utilisation pratique de l'approche basée sur les arguments développée pour analyser les essais cliniques. Ce système de démonstration a été intégré dans le contexte du projet Covid-on-the-Web pour créer des données liées riches et exploitables sur le Covid-19.

Mots clés: traitement automatique du langage naturel, extraction d'information, fouille d'arguments

Abstract

In the latest years, the healthcare domain has seen an increasing interest in the definition of *intelligent* systems to support clinicians in their everyday tasks and activities. Among others, this includes novel systems for the field of Evidence-based Medicine. The latter relies on the principle of critically appraising medical evidence and combining high quality evidence with the individual clinical experience of the practitioner with respect to the circumstances of a patient to achieve the best possible outcome. Hence, most of the proposed *intelligent* systems aim either at extracting information concerning the quality of evidence from clinical trials, clinical guidelines, or electronic health records, or assist in the decision making processes, based on reasoning frameworks. The work in this thesis goes beyond the state-of-the-art of currently proposed information extraction systems. It employs Argument Mining methods to extract and classify argumentative components (i.e., evidence and claims of a clinical trial) and their relations (i.e., support, attack). An Argument Mining pipeline is proposed and further enhanced to integrate additional information inspired by prevalent biomedical frameworks for the analysis of clinical trials. These extensions comprise the detection of PICO elements and an outcome analysis module to identify and classify the effects (i.e., improved, increased, decreased, no difference, no occurrence) of an intervention on the outcome of the trial. In this context, a dataset, composed of 660 Randomized Controlled Trial abstracts from the MEDLINE database, was annotated, leading to a labeled dataset with 4198 argument components, 2601 argument relations, and 3351 outcomes on five different diseases (i.e., *neoplasm*, *glaucoma*, *hepatitis*, *diabetes*, *hypertension*). Various Machine Learning approaches ranging from feature-based SVMs to recent neural architectures have been experimented with, where deep bidirectional transformers obtain a macro F_1-score of .87 for argument component detection and .68 for argument relation prediction, outperforming current state-of-the-art Argument Mining systems. Additionally, a Proof-of-Concept system, called ACTA, was developed to demonstrate the practical use of the developed argument-based approach to analyse clinical trials. This demo system was further integrated in the context of the Covid-on-the-Web project to create rich and actionable Linked Data about the Covid-19.

Keywords: Natural Language Processing, Information Extraction, Argument Mining

Acknowledgements

First and foremost, I wish to express my deepest gratitude to my PhD supervisors for their guidance and scientific advice, without whom this project would not have been possible. Elena and Serena, thank you for your encouraging support in difficult times, your persistent optimism, your patience and your commitment to provide me with unique opportunities throughout the years. There are many things you did, which cannot be taken for granted, for which I am grateful.

I am thankful to all members of the WIMMICS team for the intellectual exchanges and assistance. I always felt welcome and enjoyed the various collaborations. Thank you, Fabien, for fostering such a familiar atmosphere and work environment in the team. Thank you, Christine, for your assistance and patience in my struggle with the intricacies of administrative procedures. It has been a very good time in this lab, where I, as a person and researcher, could grow a lot. Merci à tous.

Special thanks to Michael, Michele and Amine. Thank you for your counsel and discussions inside as well as your friendship outside of the lab. Thanks you, Alexander, Daniel and Stefan, for your treasured friendship through the years.

Finally, I want to thank my grandma, my parents, my sister and the rest of the family for their continuous and unfailing support. Also a special thanks to you, Federica, for your love, understanding and support.

FUNDING: This work is fully funded by the French government labelled PIA program under its IDEX UCA JEDI project (ANR-15-IDEX-0001).

Contents

List of Figures

List of Tables

List of Abbreviations

ABA	**A**ssumption-**B**ased **A**rgumentation
ACTA	**A**rgumentative **C**linical **T**rial **A**nalysis
AI	**A**rtificial **I**ntelligence
AM	**A**rgument **M**ining
ARCT	**A**rgument **R**easoning **C**omprehension **T**ask
BiGRU	**Bi**-directional **G**ated **R**ecurrent **U**nit
BiLSTM	**Bi**-directional **L**ong **S**hort-**T**erm **M**emory
BOW	**B**ag-**o**f-**W**ords
CBOW	**C**onditional **B**ag-**o**f-**Words**
CNN	**C**onvolutional **N**eural **N**etwork
CORD-19	**C**OVID-19 **O**pen **R**esearch **D**ataset
CRF	**C**onditional **R**andom **F**ield
DNN	**D**eep **N**eural **N**etwork
EBM	**E**vidence-**b**ased **M**edicine
GLUE	**G**eneral **L**anguage **U**nderstanding **E**valuation
GRU	**G**ated **R**ecurrent **U**nit
HMM	**H**idden **M**arkov **M**odel
HTN	**H**yper**t**ensio**n**
IAA	**I**nter-**A**nnotator **A**greement
IRI	**I**nternationalized **R**esource **I**dentifier
KB	**K**nowledge **B**ase
LM	**L**anguage **M**odel
LOD	**L**inked **O**pen **D**ata
LSTM	**L**ong **S**hort-**T**erm **M**emory
ML	**M**achine **L**earning
MLM	**M**asked **L**anguage **M**odeling
MeSH	**Me**dical **S**ubject **H**eadings
NE	**N**amed **E**ntity
NER	**N**amed **E**ntity **R**ecognition
NLI	**N**atural **L**anguage **I**nference
NLP	**N**atural **L**anguage **P**rocessing
NLU	**N**atural **L**anguage **U**nderstanding
NN	**N**eural **N**etwork
OOV	**O**ut-**o**f-**V**ocabulary
PICO	**P**opulation **I**ntervention **C**omparison **O**utcome

PMID	**P**ub**M**ed **ID**entifier
PTK	**P**artial **T**ree **K**ernel
RCT	**R**andomized **C**ontrolled **T**rial
RDF	**R**esource **D**escription **F**ramework
RNN	**R**ecurrent **N**eural **N**etwork
RQ	**R**esearch **Q**uestion
SNLI	**S**tanford **N**atural **L**anguage **I**nference Corpus
SOTA	**S**tate-**o**f-**t**he-**A**rt
SPARQL	**S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage
SQuAD	**S**tanford **Qu**estion **A**nswering **D**ataset
SRL	**S**emantic **R**ole **L**abeling
SSTK	**S**ub**S**et **T**ree **K**ernel
SVM	**S**upport **V**ector **M**achine
tf-idf	**T**erm **F**requency-**I**nverse **D**ocument **F**requency
TK	**T**ree **K**ernel
UMLS	**U**nified **M**edical **L**anguage **S**ystem
URI	**U**niform **R**esource **I**dentifier

List of Published Papers

Tobias Mayer, Elena Cabrio, and Serena Villata, "ACTA A tool for argumentative clinical trial analysis". In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, pp. 6551–6553.

Tobias Mayer, Elena Cabrio, and Serena Villata, "Transformer-based Argument Mining for Healthcare Applications". In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020, pp. 2108–2115.

Tobias Mayer, "Enriching Language Models with Semantics". In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020, pp. 2917–2918.

Franck Michel, Fabien Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio, Olivier Corby, Raphaël Gazzotti, Alain Giboin, Santiago Marro, **Tobias Mayer**, Mathieu Simon, Serena Villata and Marco Winckler, "Covid-on-the-web: Knowledge graph and services to advance covid-19 research". In *Proceedings of the 19th International Semantic Web Conference (ISWC)*, 2020, In press.

Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata, "Generating adversarial examples for topic-dependent argument classification". In *Proceedings of the 8th International Conference on Computational Models of Argument (COMMA)*, 2020, pp. 33–44.

Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata, "Argument mining on clinical trials". In *Proceedings of the 7th International Conference on Computational Models of Argument (COMMA)*, 2018, pp. 137–148.

Tobias Mayer, Elena Cabrio, and Serena Villata, "Evidence type classification in randomized controlled trials". In *Proceedings of the 5th Workshop on Argument Mining (ArgMining@EMNLP)*, 2018, pp. 29–34.

Papers under review

Tobias Mayer, Santiago Marro, Elena Cabrio and Serena Villata, "Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials". In: *Artificial Intelligence in Medicine*. Elsevier.

Chapter 1

Introduction

This chapter explains the underlying motivation of the work presented in this thesis. It highlights the need to assist with automatic data processing tools the deliberation of clinicians in their decision making. Further, it elaborates why Argument Mining methods are good candidates to address this open challenge, especially in combination with established frameworks for evidence categorization, such as PICO. Ultimately, an overview over the structure of the thesis is given.

1.1 Background and Motivation

Clinical decision making is often intricate to a high degree. As a practitioner examining a patient, the first challenge comes with the diagnosis, the proper identification of the disease and its cause given the observed symptoms. Even though after multiple medical tests are conducted, the observed signs and symptoms might not always be sufficient to identify the exact disease. Similar symptoms can be caused by different diseases. For example, fatigue can be caused by physical exertion or emotional stress, or more serious diseases like anaemia, kidney or liver diseases and autoimmune disorders. Moreover, patients can show diverse signs of illness with different severeness for the same malady. The physician needs to consider all these factors and interpreting the symptoms to narrow down the set of potential diseases. Then, the next problem is the prediction of the course of the ailment, the prognosis, as well as potential treatments. Again, this is a highly complex case-by-case decision. The physician needs to estimate the potential effectiveness of a treatment based on his experience and the history and context of the patient. Possible adverse effects and interdependence of drugs need to be identified. Risks need to be balanced against benefits. More importantly, this all has to be done with an up-to-date knowledge of the recent research in the field. To cope with the latter, by the end of the last century, clinicians relied on narrative review articles from experts. However, the quality of these articles was volatile [1], and the results could not always be reproduced. It could not have been taken for granted that the unsystematic conclusions of these

narrative reviews were synthesized correctly [1, 2]. Thus, over the last decades, the systematic evaluation of medical evidence became more prominent. In particular, Evidence-Based Medicine (EBM) emerged to reduce the bias in reports and improve the quality of medical evidence by defining systematic evaluation standards. The core principle is a critical appraisal and judicious use of evidence, where high quality evidence is combined with the individual clinical experience of the practitioner with respect to the patient's values to achieve the best possible outcome [3]. Further, EBM should facilitate the continuing medical education of clinicians, so that they consider the results of up-to-date research in their everyday decision making. Regarding the quality of medical evidence, the focus shifted towards identifying the best available evidence in an empirical way. Clinical trials are carefully evaluated according to critical questions to determine their quality as evidence. These questions can address the experimental setup and other sources inducing bias. Moreover, the reported conclusions need to be validated and interpreted. This is what is meant by the aforementioned *critical appraisal*. With respect to the physician, who has to decide how to medicate a patient, this means that the decision has to be taken cognizant of the available evidence, where the evidence from trials or guidelines has to be compared with the circumstances of the individual patient. The latter means to decide if the evidence matches the patient's properties, and potential costs and benefits are reasonable. For this optimal healthcare, EBM should provide the required scientific framework, from the systematic evaluation of evidence to the assistance in the decision making process [1–4].

However, the EBM framework comes with some pragmatic downsides. Since most of the evidence come in the form of clinical trials, the amount of documents to process is enormous, and the manual evaluation of each trial is a laborious and time consuming task. Especially with the increasing number of published trials on the Web, it becomes challenging to effectively acquire and synthesize the available evidence [5]. Thus, forcing the clinicians' to allocate even more time for evidence search in their schedules, which are already loaded with other duties. This rises the need for systems able to support and ease clinicians' everyday activities. Accompanying the general growing popularity of Artificial Intelligence (AI), there is an increasing interest in the development of *intelligent* systems in the healthcare domain able to do exactly that: support and ease clinicians' everyday activities. These systems deal with heterogeneous kinds of data spanning from textual documents to medical images to biometrics. Concerning textual documents (e.g., clinical trials, clinical guidelines, and electronic health records), such solutions range from the automated detection of PICO[1] elements [5–7] in clinical studies to evidence-based reasoning for decision making [8–10]. These applications aim at assisting clinicians in their everyday tasks by extracting, from unstructured textual documents, the exact information they necessitate and to present this information in a structured way,

[1] Patient Problem or Population, Intervention, Comparison or Control, and Outcome.

easy to be (possibly semi-automatically) analyzed. The ultimate goal is to aid the clinician's deliberation process [11].

The aforementioned increasing amount of published data on the Web is not only limited to clinical studies. Other domains face the same challenges of analysing huge amounts of distributed and unstructured information. Together with advances in Natural Language Processing and Machine Learning, this growing problem supported the rise of a new research area called Argument Mining (AM) [12–15]. In AM the argumentation is analysed from the computational linguistics point of view, i.e., dealing with detecting, classifying and assessing the quality of argumentative structures in text. Furthermore, work in this field aims at developing approaches to aggregate, synthesize, structure, summarize, and reason about arguments in texts. Such approaches would enable users to search for particular topics and their justifications, trace through the argument (justifications for justifications and so on), as well as to systematically and formally reason about argumentation graphs. By doing so, a user would have a better, more systematic basis for making a decision. However, deep, manual analysis of texts is time-consuming, knowledge intensive, and thus unscalable. To acquire, generate, and transmit the arguments, scalable machine-based or machine-supported approaches to extract arguments are needed, which can also support argument-based decision making frameworks with machine-readable structured data. This means finding causal relationships between concepts described in a text. To address this, standard tasks in AM comprise the detection of argument components (i.e., *evidence* and *claims*) and their boundaries in unstructured text, as well as the prediction of the relations (i.e., *attack* and *support*) holding among them. As described in Chapter 9, Argument Mining methods have been applied to heterogeneous types of textual documents. However, only few approaches have applied AM methods to the medical domain [16–19], despite its natural employment in healthcare applications. As mentioned above, the reasoning stage in clinical argumentation scenarios have received considerable attention. These applications highlight the need of clinicians to be supplied with frameworks to process huge quantities of available data, as they rely on structured data as input. However, limited effort has been devoted to automatically extract this structured input from textual documents. Moreover, the demand for this kind of information cannot be directly supplied by current methods (e.g., clinical document classification [20], clinical question answering [21], or extractive summarization [22]). As explained at the beginning of this section, the medical decision making process a physician has to go through is highly complex and depends on many factors. Consequently, this forms a well-motivated need to investigate methods able to supply and support these argument-based decision making frameworks to make them practicably applicable in real scenarios. Argument mining does exactly that. It automatically detects argumentative structures, which can be the basis of Evidence-Based Medicine. For instance, the clinical trials comparing the relative merits of treatments are documents written in unstructured natural language. Thus, given its aptness to extract argumentative structures from

unstructured text, AM represents a potential valuable contribution in the healthcare domain and a powerful tool for extracting this information. Particularly, in supplementary interactions with extraction modules for other medical information, such as the automated PICO element analysis.

Hence, the goal of this PhD thesis is to start from clinical trials in natural language, define algorithms to detect and extract their argumentative structure and further aggregate it with other pertinent clinical information to provide the demanded structured data for analysing trials.

1.2 Research Questions

The road map for this project consisted of multiple stages. More precisely, each stage can be broken down into one research question (RQ) addressing different facets:

RQ1: How to adapt models from argumentation theory on large corpora of clinical text for modeling argumentation and outcome-based evidence in Randomized Controlled Trials? This question addresses the selection of a proper argumentation framework, and the adaption and extension of it for the medical domain. It further concerns the specification of which information should be extracted and how the overall output can be aggregated to be applicable.

RQ2: What computational approaches can be used to analyze arguments and evidence on Randomized Controlled Trials? This research question aims at developing a methodology for the specifications defined in RQ1. It can be further subdivided into more particular problems:

- How to define algorithms for automatically identifying arguments in medical texts? The goal is to detect components of the argument structure. This comprises challenges like the automated discrimination between argumentative and non-argumentative text units and the classification of the former into claims/conclusions and evidence/premises.
- What are suitable intra-argument relation prediction algorithms, to automatically detect the internal structure of arguments? This consists of determining how the aforementioned argument components are connected to compose the argument. In particular, it is the identification of the relations that may hold between the arguments' premises and conclusion.
- How can the argumentative structure be further aggregated with other (medical) information about the trial to make the data even more informative? This mainly addresses the integration of the PICO format, the detailed representation of observed effects of interventions and the distinction of evidence into more fine-grained labels.

RQ3: What is the impact of argumentative structures and PICO elements on evidence-based deliberation? This question investigates the practical utility of the proposed approaches and evaluates the benefit for the end user. What exactly can the developed approaches provide and what not? Where and why do they struggle? And can their weaknesses be addressed with current means?

1.3 Contributions

The main contributions of the thesis are as follows:

Contribution 1 - Creation of a New Dataset of Annotated Clinical Trial Abstracts for Argumentative Outcome Analysis First, to address RQ1, a bipolar structured argumentation model is selected [23] as the basis for the extracted information. Since the application of tools to mine arguments is very broad and given the variety of contexts where arguments can appear, a proper high-quality annotated dataset is needed, which functions as a domain ground-truth to train and evaluate automatic mining methods. With no domain specific dataset available, the creation of such a dataset was carried out by applying the aforementioned structured argumentation model to annotate a new huge resource of Randomized Controlled Trial abstracts. To the best of my knowledge, this is the largest dataset that has been annotated within the Argument Mining field on clinical data. The dataset is built from the MEDLINE database, consisting of 4198 argument components and 2601 argument relations on five different diseases (*neoplasm*, *glaucoma*, *hepatitis*, *diabetes*, *hypertension*). Furthermore, the annotations of the created dataset were extended to hold information about the effect an intervention has on an outcome and more fine-grained evidence labels. The annotation of the effects (i.e., improved, increased, decreased, no difference, no occurrence) of an intervention on 3351 outcomes is a novel aspect and an important extension for the adaption of AM on clinical trials. To foster future research in the area of Argument Mining on clinical trials, the annotation guidelines, which explain in a detailed way how the data has been annotated, and the annotated data are freely accessible. The reliability of the dataset 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.

Related Publications:

1. **Tobias Mayer**, Elena Cabrio, and Serena Villata, "Transformer-based Argument Mining for Healthcare Applications". In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020, pp. 2108–2115.

2. **Tobias Mayer**, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata, "Argument mining on clinical trials". In *Proceedings of the 7th International Conference on Computational Models of Argument (COMMA)*, 2018, pp. 137–148.

3. **Tobias Mayer**, Elena Cabrio, and Serena Villata, "Evidence type classification in randomized controlled trials". In *Proceedings of the 5th Workshop on Argument Mining (ArgMining@EMNLP)*, 2018, pp. 29–34.

4. **Tobias Mayer**, Santiago Marro, Elena Cabrio and Serena Villata, "Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials". In: *Artificial Intelligence in Medicine*. Elsevier. (under review)

Contribution 2 - Domain-specific Definition of Argument Mining Tasks and their Extensions and Evaluation for the Analysis of Clinical Trials Second, various Natural Language Processing methods are evaluated in detail on the created dataset to answer RQ2. These range from feature-based SVM approaches, which were already applied in various application domains of AM, to recurrent neural networks and other neural approach from the related work. In the end, the best performing methods rely on deep bidirectional transformers combined with task specific shallow layers to address the AM tasks of component and boundary detection and relation prediction. These architectures are further utilized to classify the effect on outcomes in clinical trials. Ultimately, a complete pipeline for processing clinical trials is proposed containing *(i)* an Argument Mining module to extract and classify argumentative components (i.e., evidence and claims of the trial) and their relations (i.e., support, attack), and *(ii)* an outcome analysis module to identify and classify the effects (i.e., improved, increased, decreased, no difference, no occurrence) of an intervention on the outcome of the trial.

Related Publications:

1. **Tobias Mayer**, Elena Cabrio, and Serena Villata, "Transformer-based Argument Mining for Healthcare Applications". In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020, pp. 2108–2115.

2. Franck Michel, Fabien Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio, Olivier Corby, Raphaël Gazzotti, Alain Giboin, Santiago Marro, **Tobias Mayer**, Mathieu Simon, Serena Villata and Marco Winckler, "Covid-on-the-web: Knowledge graph and services to advance covid-19 research". In *Proceedings of the 19th International Semantic Web Conference (ISWC)*, 2020, In press.

3. **Tobias Mayer**, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata, "Argument mining on clinical trials". In *Proceedings of the 7th International Conference on Computational Models of Argument (COMMA)*, 2018, pp. 137–148.

4. **Tobias Mayer**, Santiago Marro, Elena Cabrio and Serena Villata, "Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials". In: *Artificial Intelligence in Medicine*. Elsevier. (under review)

Contribution 3 - Practical Impact and Limitation Analysis Third, to answer RQ3, the errors of the systems are analysed in an extensive evaluation showing the issues and remaining challenges of the developed methods. Additionally, an investigation was undertaken to analyse the general robustness of the underlying bidirectional transformer model, which attests a relatively reliable handling of input with simple linguistic variations. Subsequently, general weak points of the transformer model are highlighted to indicate that this solution is still imperfect. However, its applicability to various real scenarios is demonstrated with a Proof-of-Concept system, which illustrates the impact of the argumentative information in interplay with the PICO elements. For instance, this hybrid system can identify when a claim reports an outcome as being *safe* or *efficient*, but also that the associated side effects are classified as *increased*, setting the claim into perspective. This combined analysis reveals more fine-grained categorization of the statements in RCTs.

Related Publications:

1. **Tobias Mayer**, Elena Cabrio, and Serena Villata, "ACTA A tool for argumentative clinical trial analysis". In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, pp. 6551–6553.

2. **Tobias Mayer**, Elena Cabrio, and Serena Villata, "Transformer-based Argument Mining for Healthcare Applications". In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020, pp. 2108–2115.

3. **Tobias Mayer**, "Enriching Language Models with Semantics". In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020, pp. 2917–2918.

4. Franck Michel, Fabien Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio, Olivier Corby, Raphaël Gazzotti, Alain Giboin, Santiago Marro, **Tobias Mayer**, Mathieu Simon, Serena Villata and Marco Winckler, "Covid-on-the-web: Knowledge graph and services to advance covid-19 research". In *Proceedings of the 19th International Semantic Web Conference (ISWC)*, 2020, In press.

5. **Tobias Mayer**, Santiago Marro, Elena Cabrio, and Serena Villata, "Generating adversarial examples for topic-dependent argument classification". In *Proceedings of the 8th International Conference on Computational Models of Argument (COMMA)*, 2020, pp. 33–44.

6. **Tobias Mayer**, Santiago Marro, Elena Cabrio and Serena Villata, "Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials". In: *Artificial Intelligence in Medicine*. Elsevier. (under review)

1.4 Structure

The thesis is organized as follows:

Chapter 2 describes the preliminaries, which are used throughout the thesis. It provides insights in the context and practices of the applied domain, i.e., evidence-based medicine. Further, the concepts and challenges in the research field of Argument Mining and the Natural Language Processing methods employed, which are adapted to evidence-based medicine in the context of this thesis, are presented.

Chapter 3 introduces the AbstRCT dataset which was created in the context of this thesis. After devising annotation guidelines, clinical trial abstracts were extracted via PubMed from the MEDLINE database and annotated with argumentative information. The dataset was used for almost all the experiments in this thesis. The first version of the dataset comprises mainly trials about glaucoma treatments, but also hepatitis, diabetes and hypertension subsets. In contrast, the second version consists primarily of trials about neoplasm treatments, where the aforementioned subsets from version one serve as additional test sets. The dataset comprises three annotation layers. First, annotations about argument components, such as claims and evidence. Second, argumentative relations between these components, such as attack and support, and third, Effect-on-Outcome, e.g., that an intervention increased or decreased a certain outcome.

Chapter 4 presents the Argument Mining pipeline for clinical trials I defined. The approaches addressing the two building blocks, i.e., argument component detection and relation classification, are introduced. Methods for the argument component detection include feature-based SVMs with Tree Kernels, RNNs with various word embeddings and fine-tuned transformer models. The relation classification task is addressed in two ways, i.e., as a sequence classification and multiple choice problem. For both approaches various transformer models are compared and evaluated against reference models from the AM literature. The obtained results are reported together with an in-depth error analysis.

Chapter 5 subsequently shows my work on a subtask of argument component detection, i.e., evidence type classification. A differentiation of the pieces of evidence in the dataset set is reasonable, since in EBM the results of a clinical trial are rated based on various factors. Thus, to model this variety of factors, the pieces of evidence are assigned with specialised evidence type labels, in particular, the more fine-grained label *comparative*, *significance*, *side-effect* and *other*. Various classification models including SVMs and NNs are evaluated on this task.

Chapter 6 introduces the analysis of the results of a clinical trial. Specifically, the analysis of the effect of an intervention on the observed outcome parameters. This

Effect-on-Outcome analysis is an extension of the Argument Mining pipeline, where outcomes mentioned in the argumentative components are detected and their effect subsequently classified, i.e., if an intervention has *Improved*, *Increased* or *Decreased* the outcome, or that there was *NoDifference*, or *NoOccurrence* of the outcome.

Chapter 7 is about different ways of creating linguistically simple perturbations ranging from punctuation deletion to various word-based transformations. Their impact on the robustness of current state-of-the-art Language Model based argument classification models is evaluated, with respect to both in-domain and cross-topic performance. The quality of the generated perturbations is assessed in a user study and the effect of adversarial training for argument classification is empirically evaluated. Subsequently, other known weaknesses of Language Model based transformer models, such as the ones employed in the experiments in the preceding chapters, are highlighted.

Chapter 8 demonstrates the successful applications of the proposed approaches. A Proof-of-Concept system, ACTA, shows the practical potential as a processing tool for clinical trials in general, as well as for a concrete healthcare scenario linked to the Covid-19 health emergency. ACTA is a tool for automatically analyzing clinical trial abstracts from the argumentative point of view by finding argument components and their links. Moreover, PICO elements are detected and highlighted. In the context of the Covid-19 pandemic, ACTA was updated including the extension of the relation classification and it has been employed in the Covid-on-the-Web project. Furthermore, the output is stored as RDF data through the use of ontologies for data representation. Within the project, ACTA is integrated in the overall data processing pipeline to create Linked Open Data.

Chapter 9 presents and discusses the related work in the context of evidence extraction and argumentation-based applications in Evidence-Based Medicine. It further sets the work presented in this thesis into perspective by showcasing the related development in the AM field and pointing out differences to existing approaches.

Chapter 10 concludes the thesis summarizing the contributions and remaining open questions. Furthermore, perspectives for future applications and further research directions are proposed, as well as potential plans for improvements discussed.

Chapter 2

Background

This chapter introduces the preliminaries, which are used throughout the thesis. It provides insights in the context and practices of evidence-based medicine. Further, the concepts and challenges in the research field of Argument Mining and the Natural Language Processing methods employed, which are adapted to evidence-based medicine in the context of this thesis, are presented.

In this chapter, the background in which the thesis takes place is explained. In Section 2.1, an overview over Evidence-based Medicine is given, which is the domain Argument Mining is applied to in the context of this thesis. Argument Mining itself is described in the subsequent Section 2.2, explaining the motivation, tasks and difficulties of it. The major tasks relevant for the context of the thesis are the argument component and boundary detection, and the identification of the argumentative relationships between the components. Finally, Section 2.3 gives an overview of the computational linguistic techniques applied in AM to represent natural language, ranging from early context-free representations, like bag-of-words, to recent contextualized representations from transformer models.

2.1 Evidence-based Medicine

Evidence-Based Medicine (EBM) is a practice emerged from clinical epidemiology. The core principle is a critical appraisal and judicious use of evidence, where high quality evidence is combined with the individual clinical experience of the practitioner with respect to the patient's values to achieve the best possible outcome [3].

For clinicians the continuing medical education based on new research is of utter importance. EBM was motivated by the overwhelming increment in published clinical trials and the associated difficulties of staying up-to-date in the healthcare decision making. Previously, clinicians relied on narrative review articles from experts. The quality of these articles was volatile [1] and could have been biased by a potential conflict of interest, in case where the article was published by a commercial source. Additionally, the results from these narrative reviews were not always reproducible. It could not have been taken for granted that the literature selection

was unbiased or the unsystematic conclusions were synthesized correctly [1, 2]. This *deemphasizing of the pathophysiological rationale* [24] as sufficient evidence for the decision making process created the foundations for EBM. The practitioner should be cognizant of the available evidence and carefully take decisions based on the evidence, the patient and own experience. The latter means to decide if the evidence matches the patient's properties and if the costs and benefits are reasonable. EBM should provide the required scientific framework for the systematic evaluation of evidence and assist in the decision making for an optimal healthcare [1–4]. In short, EBM consists of these five essential steps:

1. assessment of the clinical problem
2. converting the problem into answerable clinical questions
3. search for the best evidence answering these questions
4. critically appraise the found evidence for validity and usefulness and interpret what is said about the questions
5. apply the results in an appropriate manner with respect to the patient

Especially relevant for the context of this thesis are the points 2-4. For the second step, the idea is to ask well-built clinical questions which are answered by the clinical trials [25], such as questions concerning the population, intervention, comparison intervention, outcomes, time horizon or settings. These questions can be formulated in a framework called PICO, which stands for ***P****atient* or *Population*, ***I****ntervention*, ***C****omparison* or *Control*, and ***O****utcome*. Then, to prevent biased search results explicit inclusion and exclusion criteria are defined, which are search constraints that guarantee that all search results consider the inclusion criteria and not the exclusion criteria. Based on this framework and together with the defined inclusion and exclusion criteria, the literature research in step 3 is conducted. Step 4 concerns one of the fundamental concepts of EBM, the critical appraisal of the available evidence. The decisions should be taken from the best patient/population-based evidence, including an epidemiological and biostatistical analysis, such as likelihood or odds ratios and power of diagnostic information [1]. Ideally, there should be multiple trials targeting the same or similar clinical questions which are evaluated. Contrary to previous practice, where the overview of these trials was given as a narrative review without the guarantee for the unbiased systematic consideration of all the available literature, in EBM a systematic review with a subsequent meta-analysis should be conducted as a means to provide an overview. A systematic review is a structured way of processing a collection of clinical trials to limit bias. For this, evidence tables are filled, where all trials are listed in the rows and the columns describe the properties of the trials, such as the type of study, sample size, outcomes or patient demographics. This way, one can see which trials had random allocation of patients, if it was an independent blind comparison with the control group, or if patients

were assembled at the same stage of their illness. While a systematic review is a qualitative analysis, the meta-analysis is quantitative. In the meta-analysis the observed results of multiple trials targeting the same or similar clinical questions are synthesized with statistical methods. Amongst other things, this can also comprise a correction of the observed results for publication bias, which means that studies with positives outcomes tends to be published more often than negative results. In the end, there should be an assessment of the benefits, costs and potential risks for patients. Together with this general information, the patient's context and preferences, and the individual clinical expertise, the practitioner can take decisions of how to apply the available evidence with respect to potential treatments to establish the best course of action for the case under evaluation.

Originally, EBM was thought to be applied by every clinician, but correctly appraising the evidence is tedious and staying up-to-date this way is impossible. Thus, the recommendation was given to also consider EBM results from others [1]. Nevertheless, the correct and thoroughly collection and evaluation of the evidence is one of the major downsides of EBM. Reading and extracting the information from every trial has to be done manually, which is very time consuming and labour-intensive, especially with the still rapidly growing numbers of clinical trial publications [5]. Logically, many approaches have been conducted to (semi-)automatically assist in the deliberation process of the clinicians for this work. A detailed overview of these systems can be found in Chapter 9. Contrary to this previous work, in the context of this thesis Argument Mining techniques to process clinical trials were developed. The goal of using Argument Mining is not to help automatically fill evidence tables or evaluate the risk of bias, which is just one part of the evidence appraisal. Another part is scrutinizing the conclusions drawn by the authors of a trial and interpreting the results. This is where Argument Mining can assist by automatically processing the documents and creating an argumentative representation of the trial(s), which can support clinicians and practitioners in taking informed decisions.

2.2 Argument Mining

Argumentation and reasoning has become a well established field of Artificial Intelligence [26]. Argumentation is the process by which arguments are constructed, compared, evaluated in some respect, in order to establish whether any of them is warranted. While the reasoning stage and decision support has received considerable attention in the medical domain [9, 10], argument-based decision making requires structured input. Most of the time, structured data is not available, raising the necessity to develop methods to efficiently create structured arguments. One of the latest advances in artificial argumentation [27], which tackles the aforementioned problem, is the so-called *Argument(ation) Mining (AM)* [12–15]. The goal of AM is to extract and classify argumentative structures from unstructured natural language text in order to support argument-based decision making frameworks with

machine-processable structured data. This means finding causal or consequent relationships between concepts described in a text. Generally, this translates to answering the *why* question by explaining the motivations, or finding reasons and counterpoints for certain statements/propositions. Usually, the discussed topics are controversial allowing to find arguments for both sides and constructing an argumentation graph, where arguments are attacking or supporting each other.

One of the earliest work is argumentative zoning [28], where sentences are classified for their rhetorical role in a scientific paper, e.g., concerning the comparison with the scientific background or goals of the presented work. While it does not target the extraction of the argumentative structure, this classification of zones is considered a precursor for the AM area [15]. Other early seminal work comprise the detection of arguments in legal text [29, 30]. These pioneering works introducing the problem of mining arguments did not have a loud echo in the NLP community immediately. However, with technical advances overcoming previous limitations in computationally processing natural language, which are described in detail in the subsequent section, and allowing to effectively address more complex tasks such as AM, more and more attention was given to AM [14]. AM itself is a very context-dependent task, which requires deep Natural Language Understanding (NLU) and is closely related to Natural Language Inference (NLI), which lead the initial approaches to be inspired by NLI [31, 32]. Naturally, advances in Machine Learning and NLP promote the development of new AM techniques.

An AM pipeline consists of multiple tasks. The standard tasks in an AM pipeline consist in the detection of argument components, i.e., evidence and claims, and the prediction of the relations, i.e., attack and support, holding among them. The latter task can be further split into two subtasks, argument sentence detection and component boundary detection. The argument sentence detection is an anteceding step where sentences are classified as being non-argumentative or containing at least one argument component. Depending on the underlying data and use case, the more fine-grained classification into evidence and claims can be included in this step. In the case of being detected as containing arguments, the exact boundaries of the argument components, also called argumentative discourse units [12], need to be determined, since they do not necessarily span the whole sentence [33]. Both subtasks, the sentence classification and segmentation, can also be jointly tackled as one problem, which can be beneficial in some cases. With the argument components being determined, the next step in an Argument Mining pipeline is the prediction of relations holding between the components to construct the argumentative structure, i.e., which evidence supports or attacks which claims. This can be goal oriented, as in the case of argumentation scheme classification. Here, one tries to find the argumentation/rhetorical schemes the argument is composed of, where argumentation schemes are common types of reasoning patterns [34]. Contrary to that, there is the structure prediction without targeting a predefined scheme. This means that the relationship of two argument components is classified independently of the remaining

components. This relationship can also be between an argument and a general claim of a topic, where the goal is to classify if the argument component is for or against the given topic. This is similar to stance classification, where the stance of an author towards a target has to be determined [35]. While most of the proposed work only tackles certain aspects of the pipeline, the AM pipeline can also be tackled in an end-to-end manner, where both of the above described steps are solved within one (neural) architecture [36, 37].

Besides the goal of detecting the argumentative structure in an unstructured text, many nuanced subtasks has emerged over time. These tasks go beyond the pure component detection and structure prediction and aim at enriching the structure with informational features, which can be advantageous for concrete application scenarios. These subtasks can comprise argument clustering [38], argument relevance [39], argument quality [40, 41], rhetorical figure detection [42] or fine-grained evidence type classification [43].

As described in the related work (Chapter 9), Argument Mining methods have been applied to heterogeneous types of textual documents. Given its aptness to automatically detect in text argumentative structures that are at the basis of evidence-based reasoning applications, AM represents a potential valuable contribution in the healthcare domain and a powerful tool for extracting this information. Especially, in the combination with automated PICO element analysis.

2.3 Natural Language Representations

The idea of processing natural language with machines reaches back until the second half of the last century [44]. While in the beginning symbolic rule-based systems dominated the area, later statistical and Machine Learning based approaches started to play a more important role [44–46]. The goal of Machine Learning is to train a mathematical model from a collection of example data to make predictions about new data based on what was observed in the sample data. One challenge going along with this is to quantify the data, so the Machine Learning model can use it. For Natural Language Processing this is a crucial step, since human language is not encoded in numerical values. Logically, this process of creating a quantified representation of natural language is one of the oldest and most prominent problems in Machine Learning based Natural Language Processing [44, 45]. It is an essential element in all tasks, such as machine translation, natural language generation or any other classification problem [45]. Converting language into numbers poses many challenges. For example, there are many languages which are not only different in vocabulary, but also have fundamentally different structures [47]. Hence, representation models for one language might not necessarily be viable for others. Additionally, natural language is highly ambiguous and context dependent. Words can be expressions of several objects depending on the context. The meaning and connotations of a sentence depend heavily on the context and intend of the speaker [48].

Even humans speaking the same language can have difficulties understanding each other. So far, the problem to invent a model that fully understands natural language has not been solved. It remains one of the grand challenges in NLP.

In the last decade, many impactful approaches have been developed and a lot of progress was made. In the subsequent section, major approaches to tackle this problem are introduced.

2.3.1 Context-free Representations

Bag-of-words One of the simplest models to quantify natural language is the so called *bag-of-words (BOW)* model. It is a count-based model to represent text as a document-term matrix. In the base version, each sentence or document is represented as a vector, where each dimension is a word from the vocabulary, so that the total number of dimensions in the vector space equals the vocabulary size. The value of each dimension is the sum of occurrences of the word in the text unit. Depending on the text size and variety, the vocabulary/vector size can be enormous. To overcome this problem and reduce the vocabulary, the text can be pre-processed. For example, stop-words can be filtered out and inflected words can be reduced to their word stem/root (stemming) or lemma (lemmatisation). Still, for diverse text this can lead to sparse vectors, meaning that the vectors contains a lot of zeros and many of the dimensions are meaningless as part of the representation. Moreover, long text units have naturally higher count numbers of words than shorter text units. This imbalance of counts can biased the Machine Learning model.

Tf-idf One statistic to create more meaningful vectors and overcoming the text length bias is *term frequency–inverse document frequency (tf-idf)*. The idea is to weight words with respect to their occurrence in the text unit (term frequency) and overall occurrence in the corpus/collection of documents (inverse document frequency). The term frequency can be normalized and therefore make the measure text length independent, by dividing it by the number of total words in a text unit. While the term frequency is similarly representative to the bag-of-words model, the inverse document frequency lowers the weights for words which occur often across all documents and are therefore less representative for a single text unit. Tf-idf is also a common technique to filter out stop-words by setting a threshold.

N-grams So far, only single words have been considered independent of their surroundings. However, as said before, context plays an important role. While it is hard to model larger context with just counting occurrences – techniques integrating context are described in the following section – direct neighbouring words can be a valuable information which can be included on a count basis. Instead of counting single word occurrences one counts co-occurring word combinations, which can be useful to disambiguate words or detect negations and model their effect on the

change of polarity. While, these word combinations, *n-grams*, can be set to an arbitrary length n, the most common are bi- and trigrams, i.e., combinations of two and three words. Usually the necessary information to identify a collocation, i.e., *New York*, or a negative context are within the range of the two neighbouring words. Longer n-grams can, again, lead to sparse representations of vectors and promote the out-of-vocabulary (OOV) problem, where n-grams occurring in the test set have not been seeing during training, ergo are not in the vocabulary, and thus have to be ignored. Similarly to the bag-of-word, which can be interpreted as a unigram model, tf-idf can be applied to the n-grams. A common technique to further reduce the vector size is feature extraction. Here, the most representative dimensions important as features of discrimination for the Machine Learning model are considered. This optimization problem can be either solved following a heuristic search or trial-and-error strategy. Furthermore, n-grams can be used as a statistical Language Model (LM), by estimating the probability of a word given the previous context. For simplicity reasons, one assumes the Markov property. This means that a word at position n only depends on the last word at position $n-1$.

Word Embeddings Previous techniques result in a *one hot vector* representation of a word, where all dimensions but one are zero. Only text units longer than one word have vectors with more than one value. Additionally, the problem of a large feature space remains, even after feature extraction. For neural networks (NN), the use of these representations are computationally expensive and inefficient. NNs take single word vectors as input. Especially for sequence modelling, one hot vector representations are highly inefficient, since they are huge and do not contain a lot of information. A solution to this problem are *word embeddings*. The idea behind is to model each word as a dense feature vector of real numbers of size n, usually n is between 50 and 300. Each word should be represented with these n dimensions. This means that the Machine Learning model should learn to capture as many properties of a word as possible within these fixed feature dimensions. For this reason, contrary to count-based models where the dimensions are defined by the vocabulary, the dimensions of word embeddings are not interpretable by humans. This dense representation provides the capability to integrate the polysemy of words, which are defined by the context. Also, words with similar meanings should result in similar vectors[1], since they share properties. For example, the vectors of *France* and *Paris* should be in a comparable relation to each other like the vectors of *Italy* and *Rome*, or *Japan* and *Tokio*.

Static word embeddings are pre-trained with a neural network and are used as a look-up table for word features for the actual Machine Learning model, which is trained to solve a task specific problem. It has the advantage that the word embeddings do not need to be trained on the task specific dataset, which might be significantly smaller, but come from a more general and representative collection of texts.

[1]The distance is usually measured as the cosine similarity between vectors.

Word embeddings can also be pre-trained on domain specific data, such as the medical domain, to better capture the specificities of the desired domain. In general, the pre-training is conducted on a huge corpus of text. The tasks, the embedding model has to solve to learn the word representations, can differ depending on the type of word embedding. Exemplary, one of the first approaches is *word2vec* [49], where two tasks are proposed: (i) conditional bag-of-words (CBOW) and (ii) skip-gram prediction. In the latter case, word vectors are learned by predicting context words given a target word. For CBOW, the task is to correctly predict a target word given context words in a previously defined window size. Word2vec has the disadvantage that it only considers the local properties of the context within the given window size during training. Subsequent approaches try to also integrate more global information in the learning process. One example is Global Vectors or *GloVe* [50], which is based on aggregated global word word co-occurrence statistics. Similarly to n-grams, both of the aforementioned word embeddings suffer from the out-of-vocabulary problem. While domain specific pre-training with a specialised vocabulary can be beneficial, it does not solve the problem entirely. A different approach to this problem is to compose the vocabulary of smaller units, i.e., sub-word tokens, as proposed in [51].

In this thesis, it was experimented with various types of these word embeddings as input representations for NNs and other ML models, each with their own advantages and disadvantages. The full description of the used embedding types, their advantages and singular properties can be found in Section 4.1.2.

2.3.2 Contextualized Representations

Embeddings can be used as word representation features, which serve as the input for Machine Learning models. They capture the meaning of words better than count-based approaches. While they do integrate the different readings of a polysemic word in pre-training, when they are applied, the surrounding context of each word is not considered. A word has always the same general word vector independent of the actual meaning in a specific context. That is why they are also called *static* embeddings. But context is essential for the intended and perceived meaning of a word. In the case of static embeddings, the hypothesis is that a vector encodes ideally all meaning variants of a word and the surrounding words with their vectors push the representation in the vector space in the direction of the correct meaning. This principle delegates the word sense disambiguation to the Machine Learning model which uses the word embeddings as input for the classification task, e.g., a convolutional or recurrent neural network. This is an extra task the Machine Learning model has to solve besides its main task, allocating parts of the model capacity. Ideally, the model capacities should be used for the main classification task and not for word sense disambiguation. To this end, context-aware embeddings are needed, which select the right meaning of a polysemic word and shift the vector in the corresponding direction before it is fed to the classification model. These dynamic embedding types which take the full or partial context into account are called *contextualized*

embeddings.

ELMo One of the first and most prominent models for contextualized embeddings is *Embeddings from Language Models (ELMo)* [52]. It consists of a bi-directional Language Model meaning that the LM takes the preceding and succeeding context words into account. The bi-directional LM consists of two uni-directional Language Models. First, a forward Language Model which processes the input from left to right, and second, a backward Language Model which processes the input from right to left. Both LMs consist of stacked Long Short-Term Memory (LSTM) [53] neural networks, which are a subtype of Recurrent Neural Networks (RNN) with a memory cell to capture long range dependencies. They are trained separately on enormous amount of text, both on the same Language Modelling task, which was already used for statistical Language Modelling, i.e., predicting the next word given the current word. The two uni-directional output representations are concatenated to form the final bi-directional representation, where each token has its own representation which was created dynamically depending on the surrounding words. The pre-trained network can be added to any other neural network as a text encoding block. Also, ELMo eschews the out-of-vocabulary problem, since it is a character-based model. A similar approach is described in [54], where word representations are concatenated vectors of hidden states in a character-based bi-directional RNN Language Model.

ULMFiT While ELMo was a first step towards contextualizing embeddings, the pre-trained weights are only used to get the embedding of a word. The model for the downstream classification task itself has still to be trained from scratch. In computer vision, the concept of *transfer learning* was successfully applied, where a model learns a specific task and then transfers and reuses this knowledge on other tasks. For example, a model learns general features on a huge image dataset like ImageNet during pre-training and is then fine-tuned on a domain specific smaller dataset with specialized images for a certain task. For NLP, this means to leverage the weights of a Language Model learned during pre-training for a downstream task. A fine-tuning technique for efficient transfer learning was first introduced with *Universal Language Model Fine-tuning for Text Classification (ULMFiT)* [55]. This next evolution of training neural networks overcomes the problem of training models from scratch. The fundamental assumption is that different layers in a model capture features with different granularity. The lowest layer, for example, captures the most general information. In the case of language, this general property is the syntactic structure, while higher layers capture then more semantic related features. This means that ideally each layer should have an adapted learning rate corresponding to the sensitivity of the model to learn this feature. This distributed encoding of features is then leveraged in fine-tuning the model, similar to image processing. The existing problem up to this point was that with fine-tuning the pre-trained weights were changed

too drastically, throwing away the information learned during pre-training. The proposed solution in ULMFiT for this was gradual unfreezing. This means that the fine-tuning becomes an iterative process of unfreezing one layer after another. So, in the first step the last layer of the model is unfrozen and the model fine-tuned for one epoch. In the subsequent step, the next lower layer is unfrozen and the model is further fine-tuned for one epoch. This process is repeated until convergence is reached. In this way, the model has more training epochs for the more difficult and task specific features to learn, which are encoded in the higher layers of the model. The low level features like syntactic structure, which representation should be task unspecific, are modified in fewer epochs, since of their general validity the weights should only be changed slightly.

Transformer So far, the presented sequence models are based on recurrent neural architectures. RNNs come with certain drawbacks. For example, it is hard to capture long range dependency relations, because the signal has to be passed through numerous operations/time steps until it reaches the target state to encode the dependency [56]. This sequential nature not only prevents computational parallelization, but also hinders transmitting the signal, which is stronger diminished the longer it is passed through the network. Countering this with more hidden dimensions is not feasible due to memory constraints, which processing longer sequences would exceed. An impactful approach to model sequences without any recurrent or convolutional architectures, solely relying on attention mechanisms, is the *transformer* model [57]. This is a stacked encoder-decoder structure, which draws global dependencies only with multi-head attention allowing the decoder to attend to different words simultaneously for each token. Formally, the transformer is a combination of an encoder, which maps an input sequence $(x_1, ..., x_n)$ into a hidden representation, and a decoder, which translates the hidden representation into a target sequence $(y_1, ..., y_m)$. The encoder consists of N stacked layers, where each layer consists of two sub-layers. The first layer is a multi-head self-attention layer, which gets concatenated WordPiece token embeddings [58] and positional embeddings of the input sequence. The second layer is a fully-connected dense layer. Each layer is surrounded by a residual connection, and the output of the sub-layer is layer normalized. The attention layer employs Scaled Dot-Product Attention [57], where each attention function for a set of queries and key, value pairs is projected A-times in parallel. The decoder consists of the same layers as the encoder plus one extra multi-head attention layer for the output of the encoder. Since the attention mechanism passes all hidden states from the stack of encoders, the decoder can focus on multiple parts of the input sequence for each processed token. The decoder embeddings are shifted by one position, and the attention layer is masked to only attend to previous positions. The architecture is illustrated in Figure 2.1

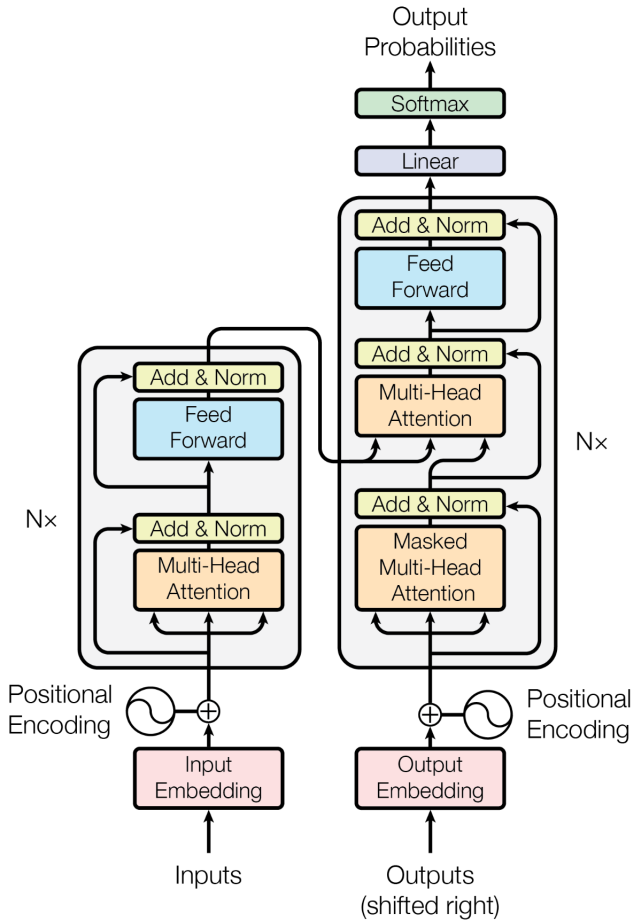

FIGURE 2.1: The transformer model architecture. Figure drawn from [57].

BERT A significant milestone in NLP founded on the aforementioned concepts and ideas of contextual representations and transfer learning is *Bidirectional Encoder Representations from Transformers (BERT)* [59]. BERT made significant improvements over state-of-the-art results on 11 NLP tasks, becoming the temporary leader spearheading the leaderboard of the General Language Understanding Evaluation (GLUE) benchmark [60] and pushing the top scores substantially. The characteristic of BERT is that it is a deeply bi-directional transformer architecture, while previous transformer models, like the OpenAI Generative Pre-trained Transformer (GPT) [61] employ a uni-directional left-to-right architecture. The idea of bi-directionality has been existing before, but it was first successfully applied to transformers with BERT. Contrary to the well established Language Model pre-training task of next word prediction, which is necessarily uni-directional, the authors propose a new task, which is capable of taking both directions into account. The *masked Language Model (MLM)* objective is inspired by the Cloze task [62], where one random word in the input is masked and the objective is to predict the original masked word. But unlike the next word prediction task, the MLM objective enables the representation to take advantage of both, the left and the right, context. This is quite similar to the optimization problem of the CBOW version of Word2vec with the difference that the MLM is not

limited by a context window size and takes the full context into account. Additonally, BERT is pre-trained with a second task, i.e., the *Next Sentence Prediction*. This is motivated with the fact that many NLP downstream tasks, such as Natural Language Inference[2], test the capability of a model to capture the relationship between two sentences. Here, as the name lets suggest, the task is to classify two given sentences as following each other or not. Based on the idea of transfer learning, BERT can be fine-tuned on downstream tasks in an end-to-end manner, which is relatively inexpensive compared to the pre-training phase.

Many succeeding work is based on the core concepts of BERT. They re-train BERT on domain specific corpora, such as SciBERT [63] and BioBERT [64], other languages [65] or cross-lingual data [66]. Other approaches modify the pre-training procedure to outperform BERT on the GLUE tasks or adapt BERT for other tasks. For example, the authors of RoBERTa [67] exchange static with dynamic masking, use larger byte-pair encoding and batches size, and increase the size of the dataset. In SpanBERT [68], the MLM is extended to mask consecutive words (spans). Both approaches show improved performance compared with the original BERT. Furthermore, there are approaches infusing more linguistic [69–71] or other domain knowledge, i.e., *Enhanced language RepresentatioN with Informative Entities (ERNIE)* [72], which includes knowledge graphs into the pre-training. Here, the MLM is complemented with an entity masking task.

In general, a whole new research field evolved. This *BERTology* investigates what and how the attention-based transformer models are actually learning [73, 74] or how to compress them efficiently [75–77] to save resources during pre-training. Chapter 7 elaborates more on the robustness and advanced pre-training techniques of these models. However, developing a new model is more than pre-training. It requires to find a good architecture and hyper-parameters as well, meaning that the model has to be trained and re-trained multiple times, which is an extremely computational expensive and power-gulping procedure [78].

[2]Formerly known as textual entailment.

Chapter 3

Creation of the AbstRCT Dataset

This chapter introduces the AbstRCT dataset which was created in the context of this thesis. After devising annotation guidelines, clinical trial abstracts were extracted via PubMed from the MEDLINE database and annotated with argumentative information. The dataset was used for almost all the experiments in this thesis. The first version of the dataset comprises mainly trials about glaucoma treatments, but also hepatitis, diabetes and hypertension subsets. In contrast, the second version consists primarily of trials about neoplasm treatments, where the aforementioned subsets from version one serve as additional test sets. The dataset comprises three annotation layers. First, annotations about argument components, such as claims and evidence. Second, argumentative relations between these components, such as attack and support, and third, Effect-on-Outcome, e.g., that an intervention increased or decreased a certain outcome. This chapter describes the results published at the International Conference on Computational Models of Argument (COMMA-2018) [79] and the European Conference on Artificial Intelligence (ECAI-2020) [80].

To address AM on clinical trials as a supervised classification problem and experiment with various approaches to extract argumentative information, annotated examples are required on which the classifier can be trained and evaluated. However, for AM in the healthcare domain no annotated dataset was available. Thus, early work of this thesis addressed this problem and covered this gap by creating a first version of a new annotated dataset of Randomized Clinical Trial abstracts, with annotations for the different argument components (evidence and claims), the AbstRCT dataset. The first version of the dataset, *AbstRCT v1*, with coarse labels contained 919 argument components (615 evidence and 304 claims) from 169 abstracts comprising 4 different diseases, i.e., *glaucoma*, *hypertension*, *hepatitis b* and *diabetes*. The first line of SVM-based experiments was conducted on this collection of trial abstracts, see Chapters 4.1.1 and 5. Consecutive experiments with neural architectures on AbstRCT v1 showed less promising results, which was attributed to the data hunger of neural networks and the relatively small size of the dataset. Therefore, in a second annotation phase, 500 additional trial abstracts were collected, annotated and

added to the dataset. Moreover, an important part of the argument structure, i.e., the relation annotation, was missing in the first version of the dataset and thus added in this second annotation phase. With the addition of the 500 trials, the possibility of changing the main topic to a more body-part-unspecific disease arose, details are described in Section 3.1. This was motivated by the necessity to make robust predictions about the generalizabilty of a model. This way, the topics from AbstRCT v1 could be reused as smaller disease specific subsets functioning as separated test sets to examine the potential generalizabilty of a trained model. To this newer dataset with 4198 components and 2601 relations in total is referred to as *AbstRCT v2*.

Furthermore, after collecting feedback from medical domain experts, I decided to incorporate information about the observed outcome in the argument structure. I expected that this additional information makes the argumentative approach to clinical trials more approachable for medical personnel, which usually does not have any background in argumentation theory, but is very familiar with the meaning and use of PICO elements. Thus, AbstRCT v2 was annotated in a third phase with Effect-on-Outcome information. In total, the AbstRCT v2 dataset is composed of the following three types of annotations:

- **Argument Components**: Comprising major claims, claims and evidence, where a *major claim* is a general statement about properties of treatments or diseases, a *claim* a concluding statement, and an *evidence/premise* an observation or measurement in the study.
- **Argumentative Relations**: The relations are connecting argumentative components to form the graph like structure of an argument. Components can be either *supporting*, *attacking* or *partially-attacking* other components.
- **Effect-on-Outcome**: Describes the effect an intervention has on each outcome (evaluated parameter) of a study. Effects were annotated when they *improved*, *increased*, or *decreased*, or when there was no difference observable or an outcome did not occur.

In the following section, the type of data, i.e., Randomized Controlled Trials, and the collection process is described. Subsequently in Section 3.2, an detailed overview of the various annotations and phases is given. The inter-annotator agreement (IAA) for all tasks and cases of disagreement are presented in Section 3.3. Finally in Section 3.4, the statistics about both versions of the AbstRCT dataset are detailed.

3.1 Data

In this section, the underlying data contained in the dataset is presented. In particular, the first part introduces the type of data which was used, i.e., Randomized Controlled Trials, and gives an understanding of why it was chosen. Subsequently,

the single phases of the data collection are explained in detail and the specification for both versions of the dataset are given.

3.1.1 Type of Data

To be in line with EBM guidelines, Randomized Controlled Trials (RCT) were chosen to be the study types, which would be annotated for creating the dataset. In particular, I decided to restrict the annotations to the abstracts of the trials following the argumentation of [5], that *"abstracts are the first section readers look at when evaluating a trial"*. Also, in related work experiments were limited to trial abstracts, because practically for literature search, medical researcher just skim through the abstract in order to evaluate if a study matches the criteria of interest [81]. Moreover, abstracts are freely accessible, while full text articles may require a paid subscription to unlock. The documentation of the study is defined by the CONSORT[1] policies, which guarantee that all necessary information of a clinical trial is stated in the abstract of the published paper. More specifically, the abstract is structured with multiple labels: *background*, *objective*, *methods*, *results* or *conclusion*. The publication policies ensure a minimum consensus of provided information, which makes the studies comparable and ideal for building a dataset.

As stated in Chapter 2.1, EBM builds the decision-making on analysing scientific information from systematic reviews of clinical trials. While clinical trials also comprise observational studies, in EBM one opts for Randomized Controlled Trials, which provide more compelling evidence [82] than the observational studies, making RCTs the most valuable sources of evidence for the practice of medicine [83]. Albeit there are more factors for this decision, one crucial aspect is the random process of assigning trial participants to at least two comparison groups, which eliminates selection bias. This randomized allocation of participants allows the use of probability theory, the likelihood that any difference between the groups was by chance can be estimated [84] and further exploited in statistical meta-analyses. Generally in a RCT, one group receives the intervention under assessment, while the other group, the control group/arm, receives either an established treatment, a placebo or no intervention at all. The intervention efficacy is determined as a comparison with respect to the control group(s). Caused by this comparative nature of the underlying data, for AM, this means that the argumentation is also build mostly on relative statements. Concretely, in the AbstRCT dataset about 70% of the annotated argumentative components contain either an explicitly stated comparison or an implicit comparison reported as measured values.

[1] http://www.consort-statement.org/

3.1.2 Data Collection

AbstRCT v1

For building the first version of the dataset, the same abstracts were selected, which were used in the dataset of glaucoma RCT abstracts of Trenta et al. [5]. This dataset is annotated with PICO elements. Trenta et al. [5] retrieved the RCT abstracts directly from PubMed[2] using three search strategies: (Strategy 1) titles or abstracts containing the word *"Glaucoma"* and that specified that the studies were randomized clinical trials; (Strategy 2) titles containing at least one element of a list of prescription drugs recognized as those used typically in the treatment of glaucoma or ocular hypertension and that specified that the studies were randomized clinical trials; and (Strategy 3) titles containing at least one element of a list of surgery procedures, identified as those typically used in the treatment of glaucoma or ocular hypertension, and that specified that the studies were randomized clinical trials. Given that in such work the authors' goal is different from the primary goals of the thesis – they extract the above mentioned PICO elements: patient group, intervention and control arms, outcome measure description and its measurements in the two arms – a new annotation process for the Argument Mining task is carried out on the same data. Moreover, given that one goal is to show the portability of the system to RCT abstracts on different diseases, 60 additional abstracts on diabetes, hepatitis and hypertension were extracted from PubMed, following Strategy 1 in [5].

AbstRCT v2

The second version strongly relies on and extends the previous version of the dataset. To obtain more training data, 500 additional abstracts were extracted with PubMed following the same aforementioned strategy. Contrary to AbstRCT v1, for AbstRCT v2 *neoplasm* was selected as a topic, assuming that the abstracts would cover experiments over dysfunctions related to different parts of the human body (providing therefore a good generalization as for training instances). And indeed, the found trials cover various different types of neoplasm, with breast and lung cancer being the most prominent types throughout the collected trials. Neoplasm as such can be either benign or malignant, but the vast majority of articles is about malignant neoplasm (cancer). In the context of this thesis, it is still referred to as *neoplasm*, since this was the MeSH[3] term used for the PubMed query.

[2]PubMed (`https://www.ncbi.nlm.nih.gov/pubmed/`) is a free search engine accessing primarily the MEDLINE database on life sciences and biomedical topics.

[3]MeSH is a controlled vocabulary thesaurus used for indexing articles in life sciences.

3.2 Annotation

In all stages of the dataset creation, annotation was started after a training phase, where amongst others the component and outcome boundaries were topic of discussion. Gold labels were set after a reconciliation phase, during which the annotators tried to reach an agreement. While the number of annotators vary for the three annotation phases (component, relation and Effect-on-Outcome annotation), the inter-annotator agreement was always calculated with three annotators based on a shared subset of the data. The third annotator was participating in each training and reconciliation phase as well.

In the following, the data annotation process of the argument component layer conducted for AbstRCT v1 and v2, the argumentative relation layer for the whole dataset of AbstRCT v2, and the Effect-on-Outcome layer also on the whole AbstRCT v2 is described. More details can be found in the annotation guidelines, which were released with each version of the dataset. The guidelines defined in [85] for Argument Mining annotation on persuasive essays serve as a basis for the development of the AbstRCT annotation guidelines, which are adapted to the clinical trial scenario. The underlying assumption of the guidelines is that a *statement* or *claim* is an assertion that deserves attention [86]. Consequently, to validate if a certain *claim* holds under specific conditions, one needs *evidence* either supporting or attacking that claim. The guidelines are available together with the AbstRCT v2 dataset here: `https://gitlab.com/tomaye/abstrct`.

3.2.1 Argument Components

The argument components as a whole are divided into three parts, one of which are claims, major claims another, and those which validates their conditions are called *premises*, or *evidence* in the following. In the successive sections, example annotations of the abstract or parts of it are shown, where **claims** are written in bold, major claims are highlighted with a dashed underline, and *evidence* are written in italics. An illustration of an annotated abstract is shown in Example 3.2.1. Two annotators with background in computational linguistics[4] carried out the annotation of the 500 abstracts on neoplasm, while the components in AbstRCT v1 were annotated by three annotators.

Example 3.2.1 Extracellular adenosine 5'-triphosphate (ATP) is involved in the regulation of a variety of biologic processes, including neurotransmission, muscle contraction, and liver glucose metabolism, via purinergic receptors. [In nonrandomized studies involving patients with different tumor types including non-small-cell lung cancer (NSCLC), ATP infusion appeared to inhibit loss of weight and deterioration

[4] In [18], researchers with different backgrounds (biology, computer science, argumentation pedagogy, and BioNLP) have annotated medical data for an AM task, showing to perform equally well despite their backgrounds.

of quality of life (QOL) and performance status]. We conducted a randomized clinical trial to evaluate the effects of ATP in patients with advanced NSCLC (stage IIIB or IV). [...] Fifty-eight patients were randomly assigned to receive either 10 intravenous 30-hour ATP infusions, with the infusions given at 2- to 4-week intervals, or no ATP. Outcome parameters were assessed every 4 weeks until 28 weeks. Between-group differences were tested for statistical significance by use of repeated-measures analysis, and reported P values are two-sided. Twenty-eight patients were allocated to receive ATP treatment and 30 received no ATP. *[Mean weight changes per 4-week period were -1.0 kg (95% confidence interval [CI]= 1.5 to -0.5) in the control group and 0.2 kg (95% CI =-0.2 to +0.6) in the ATP group (P=.002)]*$_1$. *[Serum albumin concentration declined by -1.2 g/L (95% CI=-2.0 to -0.4) per 4 weeks in the control group but remained stable (0.0g/L; 95% CI=-0.3 to +0.3) in the ATP group (P =.006)]*$_2$. *[Elbow flexor muscle strength declined by -5.5% (95% CI=-9.6% to -1.4%) per 4 weeks in the control group but remained stable (0.0%; 95% CI=-1.4% to +1.4%) in the ATP group (P=.01)]*$_3$. *A similar pattern was observed for knee extensor muscles (P =.02). [The effects of ATP on body weight, muscle strength, and albumin concentration were especially marked in cachectic patients (P=.0002, P=.0001, and P=. 0001, respectively, for ATP versus no ATP)]*$_4$. [...] This randomized trial demonstrates that **[ATP has beneficial effects on weight, muscle strength, and QOL in patients with advanced NSCLC]**$_1$.

Claims

In the context of RCT abstracts, a *claim* is a concluding statement made by the author about the outcome of the study. It generally describes the relation of a new treatment (intervention arm) with respect to existing treatments (control arm) and is derived from the described results. An example of a comparative conclusions can be seen in the Examples 3.2.2 and 3.2.3, where the latter is negated.

Example 3.2.2 [Trabeculectomy was more effective than viscocanalostomy in lowering IOP in glaucomatous eyes of white patients.]

Example 3.2.3 [Latanoprost 0.005% is not inferior (i.e., is either more or similarly effective) to timolol and produces clinically relevant IOP reductions across pediatric patients with and without PCG]

Example 3.2.4 [Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response]

Additionally to the comparative statements, *claims* can also assert general properties, e.g., that an intervention was well tolerated or had beneficial effects with respect to an outcome, like in Example 3.2.1 and 3.2.4. These statements can be in a coordinate structure, which poses the question how to split them. Ideally, the goal is to make an argument component as small and self-contained as possible. For coordinated structures, this means to split them into separated components. For

instance, in Example 3.2.4, this translates to one claim talking about the long-term ocular hypotensive effect and another one about the low rate of allergic response. Dividing the conclusions in these smaller claims makes the argumentative structure more transparent, because it is clear which assertion an evidence supports. While for a coordination it cannot necessarily be seen at first glance, especially for general outcomes with multiple aspects like *quality of life*. In practice, most of these fine-grained discrimination are prohibited by the syntactic structure of a sentence. Usually conjunctive and disjunctive coordinations are written in an elliptical manner, as it is shown in Example 3.2.4. The problem with elliptical coordinate structures is that dividing them into their single conjuncts, these conjuncts are not self-contained: the necessary contextual information, usually the omitted subject, is missing, preventing them to be a stand-alone argument component. This forces the annotators to treat them as one component increasing the complexity of the subsequent relation annotation and classification.

Major claims

Major claims are usually defined as a stance of the author in the AM literature. Here, they are defined more as a general/introductory *claim* about properties of treatments or diseases, which is supported by more specific claims. They do not necessarily occur at the end of an abstract as a final conclusion, but are mostly introduced before as a general hypothesis to be tested or as an observation of a previous study to be confirmed. A major claim with the goal of representing an introductory *claim* is shown in Example 3.2.1. Given the negligible occurrences of major claims in the AbstRCT dataset (only 3% of the components are major claims) and the structural similarity to normal claims, they are merged with claims for the classification task.

Evidence

An *evidence* in RCT abstracts is an observation or measurement in the study, which supports or attacks another argument component, usually a *claim*. Those observations comprise side effects and the measured outcome of the intervention and control arm. They are observed facts, and therefore credible without further justifications, as this is the ground truth the argumentation is based on. *Evidence* can either state exact measurements, see for instance Evidence 1-3 in Example 3.2.1, or explicitly expressed comparisons, as shown in Examples 3.2.5, 3.2.6 and 3.2.8. A common part in medical argumentation are outcomes which were not observed. For clinical decision making not only the observed change in outcomes play an important role, but also the absence of, for example, a side-effect. Section 3.2.3 elaborates more on this matter. Since these observations of absence are important, they are considered as *evidence* in the argumentation, as illustrated in Example 3.2.7.

Example 3.2.5 [*Headache, fatigue, and drowsiness were similar in the 2 groups.*]

Example 3.2.6 [*Pulse rate was significantly reduced with timolol, but not with latanoprost.*]

Example 3.2.7 [*No evidence of tachyphylaxis was seen in either group.*]

Example 3.2.8 [*Dry mouth was more common in the brimonidine-treated group than in the timolol-treated group (33.0% vs 19.4%)*]$_1$, [*but complaints of burning and stinging were more common in the timolol-treated group (41.9%) than in the brimonidine-treated patients (28.1%)*]$_2$.

Example 3.2.9 [*Mean (+/-SD) preoperative and 1-year postoperative intraocular pressures in the 5-fluorouracil group were 26.9 (+/-9.5) and 15.3 (+/-5.8)mm Hg, respectively. In the control group these were 25.9 (+/-8.1)mm Hg, and 15.8 (+/-5.1) mm Hg, respectively*]

Similarly to the aforementioned *claims*, *evidence* are often stated as conjunctive coordinations and it is important that multiple observed measures are annotated as multiple pieces of the same *evidence*. Again, the problem of how to divide them into separated self-contained units arises. In Example 3.2.5, the syntax does not allow splitting the conjunction and therefore the sentence as a whole is annotated as one single *evidence*. Exceptions can be adversative coordinations (e.g., *but, except for*). While they are usually also elliptical (see for instance Example 3.2.6), in some cases they are not and can be seen as a separated *evidence*, as illustrated in Example 3.2.8. Here, Evidence 2 is self-contained and can be processed without Evidence 1. In rare cases, *evidence* can span multiple sentences, like in Example 3.2.9. As stated before, the efficacy of an intervention in a RCT is measured as a comparison to the control group. In Example 3.2.9, each sentence on its own misses the relevant information to make the comparison from the other group. In terms of argumentation, this is a linked argument structure, where multiple premises require each other to support a conclusion. Given the interdependence of the premises in such a structure, it was decided to annotate them as one component.

3.2.2 Argumentative Relations

In order to identify complex argumentative structures in the data, it is crucial to annotate the relations, i.e., directed links connecting the components. Those relations are connecting argument components to form the argumentation graphs representing the structure of an argument. Existing approaches in AM try to form a tree structure with one root node [85]. The approach presented in this thesis is more data driven, and assumes that a trial abstract contains at least one argument in form of a tree, where an argument consists of at least one *claim* which is supported by at least one *evidence*. In practice, the average clinical trial in the AbstRCT dataset has between one and two trees, depending on the number and topic of the claims

and major claims. In general, the annotated arguments are convergent[5] or a combination of convergent and sequential[6] arguments [87]. Removing one *evidence* does not weaken the other. Given that *claims* often have a coordinate structure or make general statements, i.e., that an intervention was well tolerated, there are various independent pieces of *evidence* linked to a single *claim* making most of the arguments in the data convergent. In the AbstRCT data, sequential arguments can be seen mostly in combination with two supporting claims or major claims. There, one *claim* supported by *evidence* supports or attacks another (major) *claim*. In 19% of the cases, *claims* are linked to other (major) *claims*.

Generally speaking, an argumentative relation is a directed link from an outgoing node (i.e., the *source*) to a target node. The nature of the relation can be supporting or attacking, meaning that the source argumentative component is justifying or undermining the target argumentative component. Links can occur only between certain components: evidence can be connected to either a claim (in 92% of the cases) or another evidence (in 8% of the cases), whereas claims can only point to other claims (including major claims). The polarity of the relation (supporting or attacking) does not limit the possibility to what type of component a component can be connected. Theoretically, all types of relations are possible between the allowed combination pairs. Practically, some relations occur rather seldom compared to the frequency of others. For example, in 78% of the cases when an *evidence* is linked to another *evidence* it is an attack or a partial-attack. In rare cases, components can be unconnected. This can happen for *major claims* in the beginning of an abstract, whose function is to point out a general problem, unconnected to the outcome of the study itself.

As shown in Example 3.2.3, argument components can contain negations. For many text mining tasks negation detection and scope resolution are important subtasks, because negations entirely change the meaning of a sentence. Especially in the biomedical domain, the use of negative assertions (in particular, negating negative phrases, like *not inferior*) is abundant [88]. This poses further challenges for the automatic processing of this kind of text. In the case of AM, negations do also play an important role. Here, the impact is related rather to the correct classification of the relation than the correct linking of the components. Failing to correctly detect a negation can culminate in assigning the wrong polarity label, i.e., *attack* instead of *support*. Again, posing a great challenge for the relation classification part of the AM pipeline on clinical trials.

The annotation of argumentative relations was carried out over the whole dataset of RCT abstract in the second annotation phase, including the AbstRCT v1 subset and the newly collected abstracts on neoplasm for AbstRCT v2.

[5] A *convergent* argument consists of a *claim*, which is supported by independent *premises/evidence* [87].

[6] *Sequential* arguments consists of at least two *premises/evidence*, where one supports the other, which is supporting the final *claim*.

Attack

A component is attacking another one, if it is *(i)* contradicting the proposition of the target component, or *(ii)* undercutting its implicit assumption of significance, e.g., stating that the observed effects are not statistically significant. The latter case is shown in Example 3.2.10. Here, Evidence 1 is attacked by Evidence 2, challenging the generality of the prior observation.

Example 3.2.10 *[True acupuncture was associated with 0.8 fewer hot flashes per day than sham at 6 weeks,]*$_1$ $\xleftarrow{Attack}$ *[but the difference did not reach statistical significance (95% CI, -0.7 to 2.4; P = .3).]*$_2$

Further, an assumption is made that when the trial reports allergic reactions or other adverse effects, the author as a domain expert knows if these observations are disproportional or acceptable. So, when an intervention is claimed to be well tolerated, the *evidence* reporting these effects is considered as supporting unless the opposite is clearly stated, e.g., in form of *severe* or other modifiers.

Partial-attack

The *partial-attack* is used when the source component is not in full contradiction, but weakening the target component by constraining its proposition. Those can be implicit statements about the significance of the study outcome, which usually occur between two claims, as in Example 3.2.11. Attacks and partial-attacks are identified with a unique class for the relation classification task, because these relations are underrepresented in the dataset. In the training set only 2,5% are attack and 12% are partial-attack relations.

Example 3.2.11 **[Sentinel lymph node biopsy is an effective and well-tolerated procedure.]**$_1$ $\xleftarrow{Partial-attack}$ **[However, its safety should be confirmed by the results of larger randomized trials and meta-analyses.]**$_2$

Support

Contrary to the attack relations, the support relation is not further subdivided. While an *evidence* usually provides support for a certain aspect of the more general *claim*, it would have been often ambiguous to distinguish between partially and fully support relations, especially with respect to the impact of observed adverse effects. Thus, all statements or observations justifying the proposition of the target component are considered as supporting the target (even if they justify only parts of the target component). In Example 3.2.1, all the evidence support Claim 1. Example 3.2.12 showcases this exemplary for Evidence 3.

Example 3.2.12 *[Elbow flexor muscle strength declined by -5.5% (95% CI=-9.6% to -1.4%) per 4 weeks in the control group but remained stable (0.0%; 95% CI=-1.4% to +1.4%)*

Class	#outcomes	%
Improved	831	25
Increased	765	23
Decreased	782	23
NoDifference	897	27
NoOccurrence	76	2
Total	3351	100

TABLE 3.1: Statistics of the outcome dataset. Showing the numbers of Improved, Increased, Decreased, NoDifference and NoOccurrence classes independent of the disease-based subsets.

in the ATP group (P=.01)]$_3$ $\xrightarrow[Support]{}$ **[ATP has beneficial effects on weight, muscle strength, and QOL in patients with advanced NSCLC]**$_1$

3.2.3 Effect-on-Outcome

Argumentative structure annotations alone are for most domain-specific AM use cases sufficient. In the case of EBM, where one wants to facilitate the analysis process of trials by clinicians, further medical annotations can be beneficial. For this reason, I decided to annotate the effect an intervention has on an outcome (one of the PICO elements), e.g., if the outcome was *increased*, *decreased* or was not affected. Contrary to Lehman et al. [89], which also use these three labels[7], two extra labels are added in the here presented work, which I consider essential to fully cover the reports about an outcome. These labels are (i) the *NoOccurrence* label, when an outcome, e.g., a side effect, did not occur, and (ii) the *Improved* label for cases in which it is not clear from the text if the beneficial effect is due to a decrease or increase in the measured value of the outcome. I consider the addition of the *NoOccurrence* label important for medical argumentation, even though these reports are less frequent. For decision-making, it is not only relevant which effects were observed, but also which (side-)effects did not occur.

Note that I decided to not annotate the data with the other PICO elements. Firstly, because argumentative components contain information about the trial population only in roughly 1-2% of the cases. And secondly, there exists already a larger dataset specialised on PICO annotations [6]. Before the annotation of the Effect-on-Outcome was started, it was assessed whether the argumentative components contain enough description of those effects to have a comprehensive coverage in the AbstRCT v2 dataset. Theoretically, following the CONSORT statement [90] authors should report all PICO elements in the abstract. I found that claims contain approximately in 72% of the cases at least one PICO element (P: 2%, I/C: 51%, O: 47%) and evidence contain it approximately in 87% (P: 1%, I/C: 27%, O: 72%) of the cases. For the annotation, explicit mentions of effects on an outcome are considered. From the

[7] In my work, the *significantly* from the labels is dropped, because even though the earlier implicit assumption of significance is made, one does not know beforehand how many of the outcomes are significant, since the model cannot take components undercutting this assumption into account.

4198 argument components in AbstRCT v2, 2195 fulfilled this criteria. The others report either only the measured numerical values of outcomes (704) making the effect implicit, or general statements without an indication of a trend, e.g., that some side effect was *mild* or *common*. Moreover, many components, especially claims, give conclusive statements, e.g., that a treatment is *safe* or *efficient*, without listing the specific outcomes. Note that the annotation (and later the classification) is even more complex as about 50% of the Effect-on-Outcome containing argument components report either the outcome or the intervention in an abbreviated form. This trend is similar to the distribution of abbreviations in all argument components, where about 45% contain an abbreviation of either the intervention, or outcome or both. The detailed annotation statistic is reported in Table 3.1.

Increased/Decreased

These labels are used when it is stated that the outcome was higher, like in Example 3.2.13, or lower after an intervention, like in Example 3.2.13 and 3.2.15. Generally, it should not contain a sentiment, like *better score*. In rare cases, where an outcome was reported as *worse*, annotation guidelines were set to infer the value, e.g., a worsened side-effect usually means an increased/more intense and not a decrease occurrence. There were only a handful of cases were this was not achievable without fundamental medical expertise. These examples have been discarded.

NoDifference

An effect on an outcome is labeled as *NoDifference*, when there was no change in the outcome or when the two treatments resulted in similar values, i.e., there was no difference in the outcome between the two treatment arms. The latter case is shown in Example 3.2.13, where the *response rates* of both interventions are similar.

Example 3.2.13 Raltitrexed showed similar [response rates]$_{\text{NoDifference}}$ to the de Gramont regimen, but resulted in greater [toxicity]$_{\text{Increased}}$ and inferior [quality of life]$_{\text{Decreased}}$.

NoOccurrence

This label is used when an outcome, usually an adverse effect, was not observed, as shown in Example 3.2.14. Moreover, this example illustrates the division of coordinate structures in a single component. Contrary to argument components, the problem with ellipses preventing the division is lower, because the annotation units are smaller.

Example 3.2.14 No cases of drug-related [neutropenic fever]$_{\text{NoOccurrence}}$, [sepsis]$_{\text{NoOccurrence}}$, or [death]$_{\text{NoOccurrence}}$ occurred.

Improved

This label is used when the described outcome explicitly had a beneficial effect and no information if the measured value increased or decreased is provided, like in Example 3.2.15. There, two problems come together. First, *bleb morphology*, like *quality of life*, is a general term comprising various subscales, for instance, *bleb wall reflectivity*, *visibility of drainage route* or *presence of hyper-reflectivity area*. Second, the effect description *better* does not allow any conclusions about the measured values without concrete expert knowledge about which subscale should be increased or decreased to result in a better bleb morphology. Thus, the only certain information, which can be drawn from this statement, is that the bleb morphology improved.

Example 3.2.15 Ologen resulted in a lower long-term [postoperative IOP]$_{\text{Decreased}}$, a better [bleb morphology]$_{\text{Improved}}$, and fewer [complications]$_{\text{Decreased}}$.

3.3 Inter-Annotator Agreement

In total for all tasks, three annotators were participating in the initial annotation process. During the training phase the guidelines were refined in multiple rounds of discussion between all annotators. After the training phase, where the annotators made themselves familiar with the tasks and the data, in order to validate the annotations, the inter-rater reliability or inter-annotator agreement was calculated on a reserved and previously unseen subset of the data. The subset was sampled randomly from the collected data and each rater annotated the data independently. While the subsequent full annotation of each subtask was not always conducted with all three annotators, the corresponding IAA subset was always annotated by all three annotators and the agreement was calculated respectively.

As the statistical measure for assessing the reliability of the annotations, Fleiss' kappa [91] was used, a generalization of Scott's pi. It is suitable for a finite nominal-scale and contrary to the latter, it can be used for more than two raters. Another plausible measure would have been Krippendorff's alpha. While it is more flexible and allows other scales and missing data, the AbstRCT data is purely nominal and complete. Furthermore, having a highly imbalanced dataset could lead to instances being correctly classified by chance. Both measures control this providing a more reliable agreement score. While Krippendorff's alpha is based on the observed disagreement corrected for disagreement expected by chance, Fleiss' kappa considers the observed agreement corrected for the agreement expected by chance [92]. In the case of complete nominal data[8], both measures are similar in representing the reliability [92, 93].

[8] In the AbstRCT dataset, all N observations are assessed by all n raters, which makes the IAA subsets complete per definitionem.

Argument Components For this task, the IAA was calculated for token-level annotation. This way not only the label mismatch between *claim* and *evidence* is considered, but also the disagreement in boundary annotation. IAA among the annotators has been calculated on 30 abstracts, resulting in a Fleiss' kappa of 0.72 for argumentative components and 0.68 for the more fine-grained distinction between claims and evidence. Both values are higher than 0.61 meaning substantial agreement for both tasks [94].

Argumentative Relations Contrary to the other tasks reported in this thesis, here, the IAA was calculated not on token-level but considering each argument component as a unit. Annotation was considered as agreed, when both, the relation label and the assigned target component, were the same. IAA has been calculated on the same 30 abstracts annotated in parallel by three annotators (the same two annotators that carried out the argument component annotation, plus one additional annotator). The resulting Fleiss' kappa was 0.62, meaning substantial agreement.

Effect-on-Outcome Similarly to the argument component annotation, the agreement was calculated on token-level. Since the Effect-on-Outcome descriptions occur only on a subset of the argument components, the number of abstracts included in the IAA calculation was increased to 47. This resulted in a Fleiss' kappa of 0.81, which means almost perfect agreement [94].

3.3.1 Disagreement

In the following, the observed disagreement between the annotators and the associated difficulties, which were examined in the reconciliation phase, are discussed.

For the argument component annotation, raters disagreed on the exact determination of the boundaries. For example, conjunctive adverbs like *however* or *in general* can play an important role. In Example 3.3.1, *in general* is an important modifier which should be included in the component. Also, for phrases like *this suggests*, it can be argued that they are an important part of the argument component, because they underline the conclusive function of a claim and therefore serve as potential discriminators, in particular for cases where it is not directly clear if the statement is an observed outcome or a drawn conclusion. This is mostly the case when no exact measurement or p-value is stated, as in Example 3.3.2 for instance.

Example 3.3.1 In general, the tolerance to medication was acceptable.

Example 3.3.2 Latanoprost provided greater mean IOP reduction than did Brimonidine.

Further common disagreement was observed between claims and major claims, which can be very similar in their function as a (general) summary or conclusion.

This strengthened me in the decision to merge these two labels later in the classification. Another common conflict was the annotation of too general or co-referring components, which would be not self-contained after removing the context.

Concerning the relation annotation, most of the disagreement was not in annotating the relation label, but in assigning the target component, with an exception for the attack and partial-attack labels. As for the claims and major claims, this further endorsed the label merge for classification. Linking components lead to conflict in cases where multiple claims were very similar. One could either see a sequential structure if one considers one of the claims less specific, or two separated claims, which share parts of their evidence. In the reconciliation phase, it was decided against the latter option to avoid this kind of divergent argument structures.

For the Effect-on-Outcome annotation, one of the main disagreements between the annotators was regarding how to annotate enumerations separated by a backslash (e.g., anthralogia/myalgia); whether to annotate both as one outcome or annotate them as separated entities. It was decided to label them separately. Similar to this, the coordination of outcomes (e.g., mood, QOL or healthcare utilization) were also labeled like that, unless the separation implicates losing information related to the outcomes (e.g., liver and cardiac toxicities).

Another topic of discussion was about the inclusion of extra information/attributes relevant to the outcome or not, i.e., setting the exact boundaries. This led to further discussion on what is considered relevant information. In the end, it was decided to only include the tokens that directly affect the semantic of the outcome (e.g., *overall* QoL, *global* QoL *scores*, emphirreversible toxicity). The tokens left apart were those that do not change the semantic of such (e.g., *severity of other* toxicities, *rating of* cosmetic results, *quality adjusted* survival time). A full sentence is provided in Example 3.3.3.

Example 3.3.3 Ratings of [cosmetic results]$_{\text{Decreased}}$ decreased with time, in line with clinical observations of long-term side-effects of radiotherapy.

As previously discussed, in the dataset there are a few sentences that present two different polarities at the same time, for instance as shown in Example 3.3.4. Most of them are a comparison between the intervention and the control group where the outcome has different results for each. This was the main disagreement between the annotators, whether to annotate the outcome twice with each different result or to follow one of the group results. Ultimately, it was decided to always follow the intervention group results.

Example 3.3.4 Men in the control group had significant increases in [fatigue scores]$_{\text{NoDifference}}$ from baseline to the end of radiotherapy (P=0.013), with no significant increases observed in the exercise group (P=0.203).

Accordingly, with respect to Example 3.3.4, *control group* qualifies as the baseline and *exercise group* as intervention, meaning that the outcome *fatigue scores* is annotated as *NoDifference*. These cases pose additional challenges to the effect classifier.

Dataset	#Evi	#Claim	#MajCl	#Sup	#Att
Neoplasm	2193	993	93	1763	298
Glaucoma	404	183	7	334	33
Hepatitis	80	27	5	65	1
Diabetes	72	36	11	44	8
Hypertension	59	26	9	53	2
Total	2808	1265	125	2259	342

TABLE 3.2: Statistics of AbstRCT v2. Showing the numbers of evidence, claims, major claims, supporting and attacking relations for each disease-based subset, respectively.

3.4 Dataset Statistics

To summarize, Table 3.2 reports on the total statistics of the argumentative component and relation annotation in the AbstRCT v2 dataset. The detailed statistics of the annotated argumentative components in AbstRCT v1 can also be seen in Table 3.2, i.e., as the *Glaucoma*, *Hepatitis*, *Diabetes* and *Hypertension* subset, which come from the first version of the dataset. Table 3.1 reports on the Effect-on-Outcome annotations of the final dataset.

Concerning the argumentative annotations, there are about as half as many claims as evidence for every data split. While the average rate of evidence to claim is 2.2, the average claim has 1.87 components pointing at it. The difference is due to unconnected pieces of evidence and pieces of evidence pointing at other pieces of evidence, which are in total 22% of all snippets annotated as evidence. Major claims and attack relations are not as balanced in their distribution over the various data splits, mostly because of their rare occurrence in general. As previously stated, the average trial contains one to two argument graphs in form of trees, with the highest average of 1.98 arguments on the neoplasm subset and the lowest with 1.3 on the hypertension subset.

Chapter 4

The Argument Mining Pipeline for Clinical Trials

This chapter introduces the Argument Mining pipeline for clinical trials I defined. The approaches addressing the two building blocks, i.e., argument component detection and relation classification, are introduced. Methods for the argument component detection include SVMs with Tree Kernels, RNNs with various word embeddings and fine-tuned transformer models. The relation classification task is addressed in two ways, i.e., as a sequence classification and multiple choice problem. For both approaches various transformer models are compared and evaluated against reference models from the AM literature. The obtained results are reported together with an in-depth error analysis. This chapter comprises the work published at the International Conference on Computational Models of Argument (COMMA-2018) [79] and the European Conference on Artificial Intelligence (ECAI-2020) [80].

As stated in Chapter 2.2, the two standard tasks in AM are the argument component detection and the relation prediction/classification. While there are attempts to model these two tasks end-to-end [36], they are usually tackled separately. In the context of this thesis, a full Argument Mining pipeline is proposed, which comprises both tasks, the argument component detection and relation classification. This pipeline serves as the basis for further data augmenting extensions, as they are described in the Chapters 5 and 6. Figure 4.1 illustrates the proposed pipeline with the two major stages. The first stage is the identification of arguments within the input natural language text, in this case a RCT. This step may be further split in two different stages such as the detection of argument components (e.g., claim, evidence) and the identification of their textual boundaries. First experiments focused on the detection of arguments without paying detailed attention to the boundaries. This early work is presented in Section 4.1.1. Afterwards in Section 4.1.2, later experiments for detecting argument components, mainly based on neural networks, are showcased which scrutinize also component boundaries.

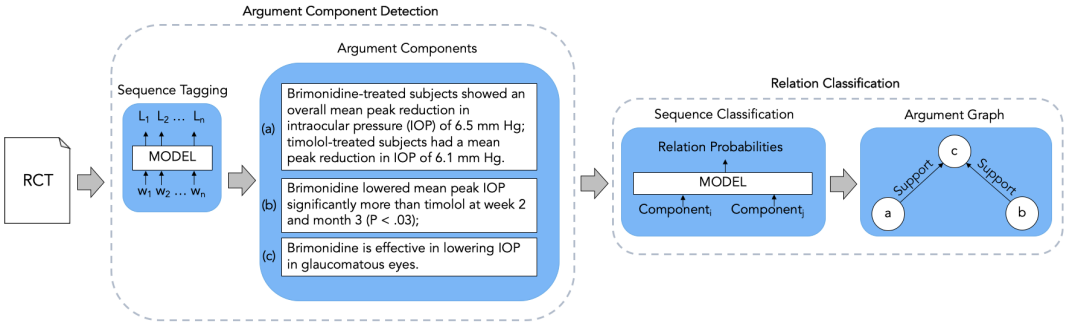

FIGURE 4.1: Illustration of the Argument Mining pipeline on clinical trials.

The second stage of the pipeline consists of predicting what are the relations, i.e., attack and support, holding between the arguments identified in the first stage. This stage is also in charge of predicting, in structured argumentation, the internal relations between the argument components, i.e., the connection between the evidence and the claim [85]. Section 4.2 explains the investigated ways to determine the argumentative structure and discussed the results of the different approaches.

4.1 Argument Component Detection

Argument component detection is typically addressed as a supervised text classification problem: given a collection of sentences, each labeled with the presence/absence of an argument component, the goal is to train a Machine Learning classifier to detect the argumentative sentences. Formally, given a dataset $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^{N}$, where x_j is a sentence and y_j is the corresponding label (whether the sentence contains an argument or not), the goal is to learn a discrimination function $f : X \rightarrow Y$ to infer the label from the input text. Such a task can be addressed by a variety of Machine Learning algorithms [13] and is evaluate in Section 4.1.1. Additionally, to include boundary detection, the argument component detection can be formulated as a sequence tagging/labeling problem. Here, x_j is not a single sentence with one label, but a sequence of tokens and y_j is the sequence of the corresponding labels. The target labels follow the BIO-tagging scheme, stating for each token that it is either the ***B**eginning*, ***I**nside*, or ***O**utside* of an argumentative component. Sequence tagging problems can be addressed with recurrent or attention-based neural networks. Representative models for both solutions are described in Section 4.1.2

4.1.1 Argument Component Detection with Tree Kernels

The method proposed in this section aims at distinguishing argumentative from non-argumentative components in natural language clinical trials and classifying the detected argumentative components into evidence and claims. As the first line of experiments conducted in the context of this thesis, I decided to rely on an existing

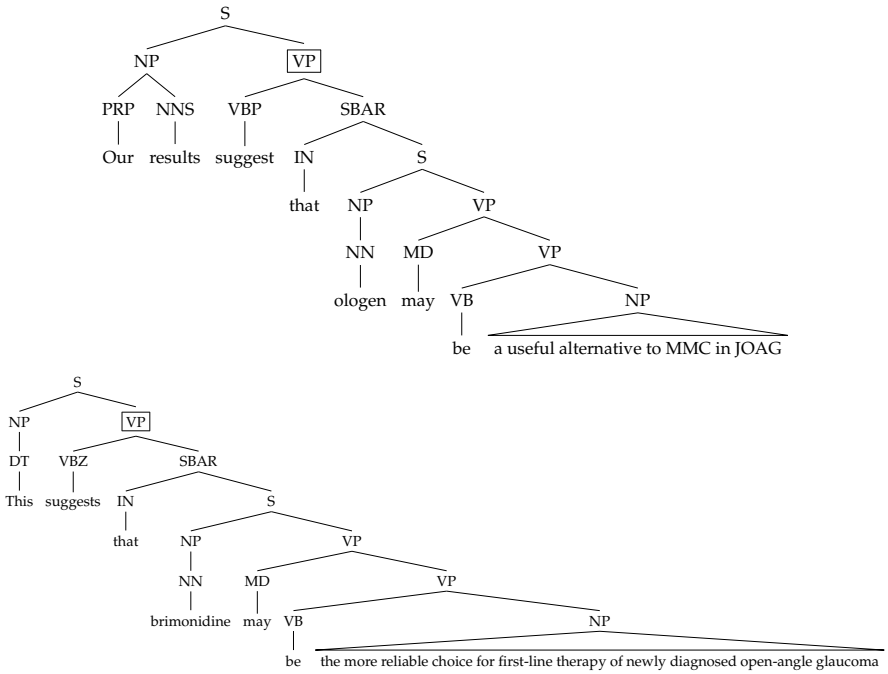

FIGURE 4.2: Constituency trees for two sentences from the corpus containing claims. Boxed nodes are common elements between the two trees.

system and tailor it to cope with the clinical data scenario. More precisely, a refined version of MARGOT [95] is provided, so that the system is able to detect evidence and claims from clinical data.

MARGOT for Clinical Trials

MARGOT[1] is an online Argument Mining server, which was designed to make Argument Mining easily accessible outside of the AM research community, and was trained on a corpus consisting of 547 Wikipedia articles [96, 97]. It was then evaluated on datasets coming from diverse genres such as persuasive essays and social media discussion threads, with encouraging results [95]. MARGOT addresses the first stage of the Argument Mining pipeline, in particular argument component detection. It carries out both claim and evidence detection, thanks to a SVM classifier that uses bag-of-words and constituency trees with subset Tree Kernels [98].

As mentioned in Section 2.3, in Natural Language Processing employing bag-of-words is a very common approach to represent sentences. This solution exploits lexical information, since each word in the vocabulary is a feature for the classifier, that is typically a Support Vector Machine. The method can be generalized to *n*-grams rather than just words. Despite its simplicity, this approach is often a strong baseline in Argument Mining [13, 95]. The methodology implemented in MARGOT

[1]MARGOT: Mining Arguments from Text. http://margot.disi.unibo.it

consists instead in a kernel machine that exploits a Tree Kernel (TK) to measure similarity between examples, namely between constituency parse trees. The key idea behind this approach is that the *structure* of a sentence is typically highly informative of the presence of an argument, or part thereof, within the sentence itself [99]. TKs aim to compare two trees by considering common *fragments*. An example of two constituency parse trees and their shared fragments is illustrated in Figure 4.2. Different definitions of fragments induce different TK functions [100].

In this line of experiments, as in the original MARGOT implementation, SubSet Tree Kernel (SSTK) [98] are employed, which offers a reasonable compromise between expressiveness and efficiency [95]. In SSTK, a fragment can be any sub-tree of the original tree, which terminates either at the level of pre-terminal symbols or at the leaves. The kernel between two trees T_x and T_z is evaluated as:

$$K(T_x, T_z) = \sum_{n_x \in N_{T_x}} \sum_{n_z \in N_{T_z}} \Delta(n_x, n_z) \tag{4.1}$$

where N_{T_x} (respectively, N_{T_z}) is the set of nodes of tree T_x (respectively, T_z), and $\Delta(n_x, n_z)$ measures the score nodes n_x and n_z, depending on the chosen definition of fragments. Given the (tree) kernel function K, the discrimination function f is defined as:

$$f(T_x) = \sum_{i=1}^{N} \alpha_i y_i K(T_{x_i}, T_x) \tag{4.2}$$

where N is the number of support vectors, and α_i is the (learned) coefficient of the i-th support vector. In our case, the problem is formulated as a binary classification task, i.e., a sentence contains an argument component or not, therefore $y_i \in Y = \{\pm 1\}$.

A very interesting characteristic of TKs is that the similarity measure implicitly allows to define a rich and expressive feature space, that basically consists of all the possible fragments that can be encountered in the parse tree.

Experimental Setup

To experiment with the proposed approach to extract argumentative information from clinical data, the first version of the AbstRCT corpus was built, with annotations for the different argument components (evidence and claims). This early version of the corpus does not comprise annotations of argumentative relations and is limited to the 169 abstracts mainly about glaucoma, for further details see Chapter 3. The model was trained on a glaucoma subset comprising 79 abstracts. From the remaining 30 abstracts about glaucoma treatments, the first test set was constructed. The remaining topics, i.e., diabetes, hepatitis and hypertension, provide three additional out of domain test sets with 20 abstracts, respectively. These four test sets get finally merged into the fifth, the *mixed*, test set.

The data was pre-processed (tokenisation and stemming), and the constituency parse tree for each sentence was computed. Furthermore, the bag-of-words features

with term frequency–inverse document frequency values were also computed. Tf-idf assigns higher weights to more distinctive words, thus lowering the impact of terms that are common among all documents, where in the case of clinical trial abstracts a document corresponds to a sentence. All the pre-processing steps were performed with Stanford CoreNLP, version 3.5.0.

Experiments were conducted with three different classifiers: (i) SSTK exploiting constituency parse trees, (ii) SVM with BOW features weighted by tf-idf, (iii) a kernel machine combining the two approaches. Two datasets were prepared to train two binary classifiers for each approach: one for claim detection, and one for evidence detection. Both training sets only differ in the labels, which were assigned to each sentence.

For tuning the hyper-parameter C (SVM regularization parameter) and the decay factor for the tree kernel, a grid search using 5-fold cross validation was executed optimizing for the F_1-score. The SVM regularization parameter C was selected from {0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0, 1.3, 1.5, 3, 5, 10, 30, 50} and the decay factor for the tree kernel from {0.1, 0.4, 0.7, 1.0}. Substituting the SSTK with a Partial Tree Kernel (PTK) [101] did not improve the results.

Results and Discussion

Evaluation of the models was conducted on multiple datasets, as described in the previous section. The binary F_1-score was computed for the tasks of (1) evidence detection, (2) claim detection, and (3) argumentative component (evidence or claim) detection. Results are shown in Table 4.1.

		Glaucoma	Diabetes	Hepatitis	HTN	Mixed
Evidence	BOW	0.84	0.79	0.74	0.80	0.80
	SSTK	**0.86**	**0.79**	**0.75**	**0.80**	**0.80**
	SSTK + BOW	**0.86**	**0.79**	**0.75**	**0.80**	**0.80**
Claim	BOW	0.75	0.68	0.62	0.64	0.65
	SSTK	**0.79**	0.73	**0.66**	**0.70**	**0.72**
	SSTK + BOW	**0.79**	**0.74**	**0.66**	**0.70**	**0.72**
Argumentative Component	BOW	0.82	0.74	0.70	0.72	0.74
	SSTK	**0.86**	**0.76**	**0.71**	**0.74**	**0.78**
	SSTK + BOW	**0.86**	**0.76**	**0.71**	**0.74**	**0.78**

TABLE 4.1: Results for the glaucoma, diabetes, hepatitis, hypertension (HTN) and mixed test set on the task of evidence, claim and argumentative component detection. Results are given in F_1 score.

The model behavior is different for claim detection and for evidence detection. As for claim detection, the best performance is still on the glaucoma set with 0.79 F_1 score and 0.75, respectively. But here, the difference between the SSTK and BOW model is significantly higher. This suggests that claims have a distinctive syntactic structure which can be learned, and that is useful to distinguish them from non-argumentative sentences and evidences. This is true also when the test set comprises

the same topic as the training set and thus the lexical approach should have a natural advantage. Furthermore, when comparing the results from the glaucoma test set with the other test sets, the performance of the SSTK does not decrease as strongly as the one of the BOW model, e.g., -0.07 vs. -0.10 F_1 score on the mixed dataset (joint test set over all domains). Differently from the case of evidence detection, here, the BOW relies more on specialized medical terminology, which differs with the individual test set domain, whereas TKs generalize better on out-of-domain data. The combined model delivers similar results compared to the pure SSTK model. Again, as for evidence detection, this suggests that the lexical information representing the characteristics of claims is also contained in the syntactic representation. For evidence detection, all models performed best on the glaucoma test set. This is intuitive, since in that case training and test domains coincide. Comparing the different models on this test set, the SSTK performed slightly better with 0.86 F_1 score, but still the difference to the BOW baseline (0.84 F_1 score) is only marginal. This difference becomes even smaller on the hepatitis dataset and vanishes completely for diabetes and hypertension. Thus, it can be concluded that evidence is not that highly distinctive with respect to syntactic structure from non-argumentative sentences or claims, while it can be easily identified by lexical information. Moreover, this lexical information is domain independent, otherwise the performance of the BOW model would strongly decrease with respect to the SSTK on the other test sets. Therefore, the distinctive vocabulary is likely to be related to the domain of statistical evaluation, rather than to medical terminology, as one could expect. These observations will need further investigation. Interestingly, the combination of the syntactic (SSTK) and lexical (BOW) approach did not increase the results, meaning that those two models share the equal amount of information representing the characteristics of evidence, and that the two models generalize equally well.

The results of the third classification problem, the detection of argumentative sentences, reflect the above described findings. The best performance for each model was obtained on the glaucoma set. The SSTK model outperforms the BOW baseline, but a combination of TKs and BOW does not increase the results. Again, the TK generalizes better over the different test set domains.

Error Analysis Comparing the outcomes of the experiments for claim and argumentative component detection tasks, the best models for the glaucoma and mixed test set perform in a comparable range. In theory, the performance for the combined task should be lower, since the claim detection has a significant lower F_1 score than the evidence detection. This can be explained when looking at the errors made by the claim classifier. A sizable amount of false positives, sentences which are classified as containing a claim, but actually do not, were sentences containing evidence. When merging together evidence and claims into argumentative components, those false positives become true positives, increasing the overall results.

1.	**predicted label:** Claim; **correct label:** NoArgument The goal of this research is to evaluate efficacy and safety of herbal medicine as compared to allopathic medicine in patients suffering from hepatitis B.
2.	**predicted label:** NoArgument; **correct label:** Claim The authors tested the hypothesis that a valsartan/cilnidipine combination would suppress the home morning blood pressure (BP) surge (HMBPS) more effectively than a valsartan/hydrochlorothiazide combination in patients with morning hypertension, defined as systolic BP (SBP) $\geq$135 mm Hg or diastolic BP $\geq$85 mm Hg assessed by a self-measuring information and communication technology-based home BP monitoring device more than three times before either combination's administration.
3.	**predicted label:** Evidence; **correct label:** NoArgument Among 426 participants (53% male, median age 35 years, median CD4 count 19 cells/μL), 31 developed hepatotoxicity (7.3%).
4.	**predicted label:** Evidence; **correct label:** NoArgument Overall, there were no significant differences in pregnancy-induced hypertension across supplement groups.
5.	**predicted label:** NoArgument; **correct label:** Evidence No patients developed additional resistance mutations throughout the study period.

TABLE 4.2: Sample classification errors for the argument component detection using SVMs with a TK.

Error analysis on claim detection indicates that a significant amount of false positives are sentences describing the objective of the RCT (Table 4.2, Example 1). This might be due to the comparative nature of the sentences, since comparative statements are common among claims. Many false negatives have complex syntactic structures (Example 2), where either the whole sentence is a claim, or it contains multiple fragments with claims. Those complex structures might have been missing from the training set.

For evidence classification, many sentences describing the participants of the studies (Example 3) have been misclassified as evidence by all approaches. This might be due to their sub-clauses containing statistical descriptions, as many pieces of evidence have too. Similarly, sentences describing the initial condition of the different groups (Example 4) were confused as evidence. The problem here is that those sentences are highly context-dependent: Example 4 could be a valid evidence, if the context was the description of the results and not the description of the initial conditions. There is no way to distinguish those cases without considering a larger context. Other misclassified evidence are negated sentences like Example 5, reporting the non-existence of an effect.

4.1.2 Component and Boundary Detection with Neural Architectures

Consecutive experiments with neural networks on the AbstRCT v1 corpus showed the necessity for a larger dataset, for more details I refer the reader to the discussion of the results in Chapter 5. To fulfill this requirement, more data was collected and

annotated, resulting in the release of AbstRCT v2 (see Chapter 3 for more details about the dataset). Almost tripling the size of the corpus, in the second line of experiments neural networks were applied for the argument component detection. Reformulating the problem as a sequence tagging/labeling task opens the potential to directly integrate and have a closer look at the boundary detection as well, i.e., which parts of a sentence are the smallest argumentative units. As I will detail in Chapter 9.2, most of the AM approaches classify the type of component assuming the boundaries of argument components as given. Thus, to merge the component classification and boundary detection into one problem, the component detection is cast as a sequence tagging task, as illustrated in Figure 4.1. Following the BIO-tagging scheme, each token should be labeled as either being at the **B**eginning, **I**nside or **O**utside of a component. As there are two component types in AM, this translates into a sequence tagging problem with five labels, i.e., *B-Claim*, *I-Claim*, *B-Evidence*, *I-Evidence* and *Outside*. To model the temporal dynamics of sequence tagging problems, usually Recurrent Neural Networks (RNN) are used. In the experiments, different combinations of RNNs are evaluated with various types of pre-trained word representations, which are introduced in the subsequent section. Each embedding method is combined with uni- or bidirectional LSTMs or GRUs with and without a CRF as a last layer. With the rise of attention-based transformer models, I modified them to suit the sequence tagging problem and overcome the common problems of recurrent architectures, such as long range dependencies. These were the first experiments on token level classification in AM by fine-tuning different transformer models.

Word Embeddings

There are two ways to create an input word representation for sequence modelling. One way is to look up the representation from pre-trained embeddings. This static method has the advantage that one does not need to train its own embeddings. However, the vocabulary is limited, and the context of the word is not considered. State-of-the-art embeddings are generated dynamically from the context of the target word based on pre-trained Language Models [52, 54, 59]. In the experiments, both kinds of embeddings are considered. Furthermore, since the AbstRCT data is from the medical domain containing very specific terminology which might not be covered in the vocabulary of general word embeddings, different approaches to overcome this problem were examined.

Static Embeddings As for the static embeddings, **GloVe** embeddings [50] are commonly used. They are based on aggregated global word-word co-occurence statistics and trained on Wikipedia and the Gigaword 5 corpus. Words are considered to be the smallest unit. In the experiments, the 100 dimensional version is used. Extended dependency-based skipgrams, short **extvec** [102] are trained also on Wikipedia, but

make use of structural information coming from dependency graphs. The embedding size is 300 dimensions. Contrary to these embeddings on word level, **fastText** [51] embeddings work on a sub-word level and are commonly used to overcome the out-of-vocabulary problem. They encode sub-word information based on a character n-gram model, and use position weights to predict words context dependent. The 300 dimensional version pre-trained on the Common Crawl and Wikipedia is used. Like fastText, Byte-Pair embeddings **BPEmb** [103] use sub-word segments to increase the capability of their vocabulary and might, because of that, be a better choice for a setting with unusual and specific terminology. Here, the segmentation is modelled with an iterative merge operation over the most frequent symbols, where a symbol is the output of the last merge operation starting on character level. They are trained on Wikipedia and embed words into a 100 dimensional vector.

Dynamic Embeddings Moving to the dynamically generated embeddings, Embeddings from Language Models (**ELMo**) [52] are generating the representation of a word by contextualizing it with the whole input sentence. They use a bi-directional LSTM to independently train a left-to-right and right-to-left character-based LM. The vectors of these models are concatenated to form a single contextualized representation of the input word. For the here presented work, the ELMo model trained on PubMed was used to have a model which is trained on the same type of data as the target data, i.e., the AbstRCT corpus. For the same reason, the on PubMed trained Contextualized String Embeddings (**FlairPM**) [54], another character-based Language Model, were used. There exists also a general one (**FlairMulti**) trained on a mix of web content, Wikipedia, subtitles and news, which was used as a direct comparison to investigate the impact of domain specific pre-training. In these embeddings, word representations are concatenated vectors of hidden states in the bidirectional RNN Language Model. A word with its context sentence is given as input into the LM. To encode the word into a contextualized representation, the forward hidden state of the last character of the word and the backward hidden state of the first character of the word are concatenated. The third dynamic embedding are Bidirectional Encoder Representations from Transformers (**BERT**) [59]. Here, the bidirectional representation is learned jointly with a transformer architecture, which will be described later in this section. The Language Model considers sub-words and position of the word in the sentence to give the final representation of a word. BERT is pre-trained on a concatenation of the BooksCorpus and English Wikipedia. For the experiments, the BERT_{base} model was used, which encodes words into a 768 dimensional vector.

Recurrent Neural Networks

As already mentioned, sequence tagging is the task of assigning a label to each token of an input sequence. To model the temporal dynamics of such sequences, Recurrent

Neural Networks are usually used. Those networks take information from past time steps into account. The sequence of vectors is processed one by one and the hidden state of a previous time step functions as the memory for the already processed sequence. Naturally, this repeated process of concatenating the whole memory state with the current state and passing in through an activation function, leads to an information loss over longer distances. That is why RNNs have only a short-term memory. To counter this effect, gates are integrated into the RNN to regulate the information flow, i.e., which information is relevant to keep and which can be discarded. The two most commonly used gated RNN architectures nowadays are the **LSTM** [53] and **GRU** [104]. The LSTM is a cell consisting of three gates (forget, input and output gate) and outputting two vectors (cell state and hidden state). The cell state routes the information flow in the LSTM, while the hidden state is used to calculate the model predictions. Concerning the various gates, the input gates determines which part of the input vector of the current step is relevant. The forget gate regulates the information with respect to what is kept in the memory. The output gate is responsible for the hidden state output. Contrary to the LSTM, the newer GRU has only two gates, i.e., a reset gate and an update gate, but builds on the same principles as the LSTM. The update gate has a similar function as the forget and input gate of the LSTM. It determines which parts of the current input are relevant and which information should be kept in the memory. The reset gate decides which information from the previous step should be forgotten. Since the hidden state of a GRU take both roles, the transfer of memory information and providing the hidden state for calculating the prediction, an output gate as for the LSTM is not required. In a direct comparison, the GRU requires fewer computations, but generally the performance of both architectures is similar and might only differ for certain tasks. For this reason, both model architectures are included in the here presented experiments.

As previously mentioned the tagging scheme used to encode the label information is the BIO-tagging scheme. This means that ideally after a B-token an I-token should follow. Modelling these constraints falls under structured prediction. Here, token-wise classification is not done independently of each other, but the full structure is predicted context dependent as a multivariate probability distribution. Usually, statistical graphical models are used to represent the distribution and to infer the most probable sequence of labels. In general, there are two families. One, generative models, such as the Hidden Markov model (HMM), which model the problem as a joint distribution $P(y, x)$. And two, discriminative models, such as Conditional Random Fields (**CRF**) [105], which model the problem as a conditional probability distribution $P(y|x)$. In the context of this thesis, CRFs are used, since they yield higher performance due to not having the need to model the distribution of $P(x)$. CRFs can be seen as a sequential extension of the Maximum Entropy model [105]. Simply speaking, they consider the predicted label of the other time steps and decode into the most probable sequence of labels. For this reason, a CRF is build on top of the RNN to enforce structured predictions in the here presented experiments.

Transformer Models

Transformer architectures have recently advanced the state-of-the-art for multiple NLP tasks [59, 61]. As described in Chapter 2.3, a transformer [57] is a combination of an encoder, which maps an input sequence $(x_1, ..., x_n)$ into a hidden representation, and a decoder, which translates the hidden representation into a target sequence $(y_1, ..., y_m)$. The encoder consists of N stacked layers, where each layer consists of two sublayers. The first layer is a multi-head self-attention layer, which gets concatenated WordPiece token embeddings [58] and positional embeddings of the input sequence. The second layer is a fully-connected dense layer. Each layer is surrounded by a residual connection, and the output of the sub-layer is layer normalized. The attention layer employs Scaled Dot-Product Attention [57], where each attention function for a set of queries and key, value pairs is projected A-times. The decoder consists of the same layers as the encoder plus one extra multi-head attention layer for the output of the encoder. The decoder embeddings are shifted by one position, and the attention layer is masked to only attend to previous positions.

Transformers can be used as features to an RNN, but also have the possibility to fine-tune the pre-trained model on a target dataset. Hence, for the experiments BERT is used feature-based as embeddings for the RNN, but also as a transfer learning model using fine-tuning. By the time of the experiments, there were already various pre-trained models available for the latter method. Beside the original BERT, which is pre-trained on the BooksCorpus and English Wikipedia, **BioBERT** [64] is pre-trained on large-scale biomedical corpora outperforming the general BERT model in representative biomedical text mining tasks. The authors initialize the weights with the original BERT model and train on PubMed abstracts and full articles. Therefore, the vocabulary is the same as for the original BERT. Contrary to that, **SciBERT** [63] is trained from scratch with an own vocabulary. While SciBERT is trained on full papers from Semantic Scholar it also contains biomedical data, but to a smaller degree than BioBERT. The uncased SciBERT model was chosen, meaning that the capitalization of words is ignored. As it was the case for the original BERT, the uncased model of SciBERT performs slightly better for sentence classification tasks than the cased model.

Originally, BERT was not designed for sequence tagging. Thus, to make it applicable for the argument component detection, I extended the transformer with task specific layers. For fine-tuning on the sequence tagging task, the hidden state representation of each word of the transformer is taken and fed into shallow layer build on top of the transformer. For this shallow layer different variants were examined. First, a dense layer mapping directly into the label space. Second, a CRF to enforce structured predictions. Third, a combination of a (bi-directional) GRU/LSTM and CRF similar to the aforementioned architecture to evaluate the various word embeddings.

Experimental Setup

For sequence tagging, each of the above mentioned embeddings were combined with either *(i)* a GRU, *(ii)* a GRU with a CRF, *(iii)* a LSTM, or *(iv)* a LSTM with a CRF. Additionally, the best performing static and dynamic embeddings were concatenated and evaluated as if they were one embedding. The *Flair* [54] PyTorch NLP framework version 0.4.1 was used for implementing the sequence tagging task. For BERT, the PyTorch implementation of huggingface[2] version 2.3 is used. Hyperparameter tuning was done with hyperopt[3] version 0.1.2. The learning rate was selected from {0.05, 0.1, 0.15, 0.2}, RNN layers {1, 2}, hidden size {32, 64, 128, 256}, dropout {0.1, 0.2, 0.5}, and batch size from {8, 16, 32}. The RNNs were trained over 100 epochs with early stopping and SGD optimizer. For fine-tuning the BERT model, the uncased base model (Bert_{base}) is employed with 12 transformer blocks, a hidden size of 768, 12 attention heads, a learning rate of 2e-5 with Adam optimizer for 3 epochs. The same configuration was used for fine-tuning Sci- and BioBERT. For SciBERT, the uncased model with the SciBERT vocabulary is used. For BioBERT, version 1.1 was selected. Batch size was 8 with a maximum sequence length of 256 sub-word tokens per input example.

The neoplasm part of the AbstRCT corpus was split such that 350 abstracts are assigned to the train, 50 to the development, and 100 to the test set. Additionally, the first version of the dataset was used to create two extra test sets, both comprising 100 abstracts. The first one includes only glaucoma, whereas the second is a mixed set with 20 abstracts of each disease in the dataset (neoplasm, glaucoma, hypertension, hepatitis and diabetes), respectively.

Results and Discussion

The results for the best performing RNN models and the best performing embedding combinations are shown in Table 4.3. Results are given on all three test sets in micro and macro multi-class F1-score and for claim and evidence, respectively. Comparing the static word embeddings, fastText with a BiGRU and a CRF is the best performing combination, where extvec is only slightly worse and is usually better for evidence classification. For the dynamic embeddings coming from LMs, the ones trained on the medical domain corpus, i.e., FlairPM and ELMo, show similar performances with a macro F1-score of .68 on the neoplasm test set. They have the edge over the non-specialized LMs like BERT with .66 or FlairMulti with .63 macro F1-score. Concatenating static and dynamic embeddings does not bring a notable difference, when taking all test sets into account.

[2] https://github.com/huggingface/transformers
[3] https://github.com/hyperopt/hyperopt

		Neoplasm				Glaucoma				Mixed			
Embedding	Model	f_1	F1	C-F1	E-F1	f_1	F1	C-F1	E-F1	f_1	F1	C-F1	E-F1
GloVe	BiGRU+CRF	.61	.58	.50	.66	.60	.52	.36	.68	.55	.50	.36	.64
extvec	BiGRU+CRF	.67	.65	.58	.72	.68	.64	.57	.72	.67	.64	.57	.71
fastText(fT)	BiGRU+CRF	.68	.66	.61	.71	.68	.65	.60	.71	.65	.60	.52	.69
BPEmb	BiLSTM+CRF	.64	.60	.59	.76	.64	.60	.52	.69	.61	.57	.48	.66
ELMo	BiLSTM+CRF	.70	.68	.59	.76	.74	.72	.67	.77	.72	.70	.67	.74
BERT	BiLSTM+CRF	.69	.66	.58	.75	.70	.68	.63	.73	.68	.66	.61	.71
FlairMulti	BiLSTM+CRF	.66	.63	.53	.72	.58	.55	.50	.60	.52	.50	.44	.56
FlairPM	BiLSTM+CRF	.70	.68	.60	.75	.74	.72	.69	.75	.70	.68	.64	.72
FlairPM + extvec	BiGRU+CRF	.68	.65	.54	.74	.74	.72	.67	.77	.68	.66	.60	.72
FlairPM + fT	BiGRU+CRF	.68	.64	.53	.75	.71	.68	.62	.74	.67	.63	.56	.71
FlairPM + BERT	BiLSTM+CRF	.70	.69	.61	.76	.71	.70	.67	.73	.68	.67	.62	.72
BERT + fT	BiLSTM+CRF	.68	.65	.55	.74	.68	.66	.60	.71	.67	.65	.58	.71
ELMo + fT	BiLSTM+CRF	.71	.68	.59	.77	.74	.72	.69	.77	.72	.70	.65	.75
fine-tuning BERT	dense layer	.82	.60	.69	.83	.77	.55	.63	.80	.80	.57	.65	.83
fine-tuning BERT	CRF	.89	.84	.78	.90	.90	.85	.81	.89	.90	.85	.79	.90
fine-tuning BERT	GRU+CRF	.89	.85	.78	.90	.89	.86	.76	.89	.90	.88	.81	.91
fine-tuning BioBERT	GRU+CRF	**.90**	.84	.87	.90	**.92**	**.91**	**.93**	.91	**.92**	**.91**	**.91**	.92
fine-tuning SciBERT	GRU+CRF	**.90**	**.87**	**.88**	**.92**	.91	.89	**.93**	.91	.91	.88	.90	**.93**

TABLE 4.3: Results of the multi-class sequence tagging task are given in micro F1 (f_1) and macro F1 (F1). The binary F1 for claims are reported as C-F1 and for evidence as E-F1.

Generally, evidence scores are higher than claim scores, leading to the conclusion that claims are more diverse than evidence. This is coherent with the findings from the previous experiments on argument component detection described in the preceding section. One explanation for this observation is that, since natural language reports of measurements in clinical trials vary mostly only in the measured parameter and its values, a distinctive lexical pattern can be learned. This assumption is further supported by the aforementioned experiments with the SVMs. There, the BOW feature was shown to carry more distinctive information for evidence than for claims, i.e., there has to be some significant evidence-specific lexical patterns. The detailed results for this are reported in Table 4.1. On the other side, claims can be made about almost everything, which reduces the number of useful lexical cues. As shown with the Tree Kernels, structural features can be a good indicator and transformer models do capture structural information in the lower layers, but apparently this is not sufficient enough to reach the detection rate of evidence.

Another observation is that the performance of the models trained on neoplasm data do not significantly decrease for test sets on other disease treatments. This fact supports the choice of a more general high level disease type like neoplasm for training the models. The performance for many model combinations even increases on the glaucoma test set. The glaucoma test set comprises only a handful of different glaucoma treatments and is therefore less diversified than the neoplasm or mixed test sets. This is ideal with respect to the application of such models, where clinicians will compare studies for a specific disease treatment. Looking at the main difference in the results, fine-tuning BERT outperforms all other model combinations, where the version with a BiGRU and CRF is the best performing model. Fine-tuning without any kind of sequence modelling on top of it results in worse performance. Especially with respect to the validity of BIO sequences, where disproportionately many invalid sequences are generated. This is not useful when extracting the components based on BIO-scheme. The direct comparison between the various options for the sequence modelling shallow layer on top of the transformer are illustrated exemplary for the BERT_{base} model in Table 4.4. The most notable difference is achieved by adding a CRF. As explained earlier, this forces the model to consider all labels of a sequence instead of making an independent prediction for each token. Interestingly, adding a uni-directional GRU or LSTM between the transformer and the CRF does not increase the overall results. One the contrary, it even lowers the performance on some test sets. Replacing the uni-directional with a bi-directional RNN increases the performance only slightly with respect to having no RNN at all. Recalling why transformers were invented, the attention mechanism is supposed to not suffer from the same problem of transmitting long distance information as it is the case for recurrent models. Interpreting the results, this means that the transformer part actually captures the necessary information for the classification task, while the sequence modelling of the RNN becomes redundant. The only marginal increase of the bi-directional GRU is most likely more due to the increase in trainable network

	Neoplasm				Glaucoma				Mixed			
	f_1	F1	C-F1	E-F1	f_1	F1	C-F1	E-F1	f_1	F1	C-F1	E-F1
dense layer	.82	.60	.69	.83	.77	.55	.63	.80	.80	.57	.65	.83
CRF	**.89**	.84	**.78**	**.90**	**.90**	.85	.81	**.89**	**.90**	.85	.79	.90
GRU+CRF	**.89**	.84	**.78**	**.90**	.88	.80	.81	.87	.89	.81	.78	.90
LSTM+CRF	.88	.65	.73	.89	.87	.63	.78	.86	.88	.64	.76	.88
BiGRU+CRF	**.89**	**.85**	**.78**	**.90**	.89	**.86**	.76	**.89**	**.90**	**.88**	**.81**	**.91**
BiLSTM+CRF	**.89**	.80	.77	.89	**.90**	.81	**.82**	.88	.89	.81	.79	.90

TABLE 4.4: Comparison of various architectures for the shallow layer extension of BERT for the sequence tagging task. Results are given in micro F1 (f_1) and macro F1 (F1). The binary F1 for claims are reported as C-F1 and for evidence as E-F1.

parameters than the actual recurrent architecture. In a direct comparison between GRU and LSTM, both RNN types deliver results in a comparable range, where the GRU does seem to show more reliable results. For example, the .65 macro F_1-score on the neoplasm test set for the uni-directional LSTM is due to the complete failure of correctly detecting *B-Claim* tokens, which the GRU counterpart does not struggle with. Similar observations were found for the bi-directional variants. Here, the BiLSTM misclassifies *B*-tokens as *I*-tokens of the correct component type. This translates into a lower macro F_1-score, because this is the average score over all labels, while the C- and E-F1 remain comparable, because they are weighted scores and the confusion of *B*- and *I*-tokens of the same type does not influence this score as strongly as the macro score.

Comparing the specialized with the general models, Bio- and SciBERT show a better performance than the general BERT model, where the cased BioBERT tends to be more reliable for the out of domain test data. This is in line with the findings that the cased transformer model works better for tasks like Named Entity Recognition (NER), which is also a sequence tagging task. The difference on the AbstRCT data is marginal: while for NER the casing of a word is relevant, for argument component detection it does not seem to be a sensitive information.

Error Analysis Despite the CRF, common mistakes for the sequence tagger are invalid BIO sequences. Especially when there are multiple components in one sentence, the tagger tends to mislabel *B*- tokens as *I*- tokens. This is due to the natural imbalance between *B*- and *I*- tokens. Training the sequence tagging without the BIO scheme using only *claim* and *evidence* as labels, poses problems when multiple components are following each other in the text. They would be extracted as one single component instead. This is a common case in concluding sentences at the end of a study, which strikingly often comprise multiple claims. Further experiments could go in the direction of weighted loss functions like focal loss to overcome this problem. Notable mistakes arise for determining the exact component boundaries. Especially in the case of connectives, e.g., *however*, which have sometimes nothing but a conjunctive function, and in other cases signal a constraint of a previous statement. Similar to the aforementioned TK-based SVM, the mistake of misclassifying

the description of the initial state of the participant groups as an observation of the study and therefore an evidence remains, see Example 4.1.1.

Example 4.1.1 (**predicted label:** Evidence; **correct label:** NoArgument)
There were no significant differences in pregnancy-induced hypertension across supplement groups.

In the study abstract these descriptions occur usually relatively close to the actual result description, which means that adding information of the position in the text will not avoid this error. While only some abstracts are structured, the full study report does usually have separated sections. This structure can be exploited when analysing full reports, and in the simplest case one would analyse only the sections of interest.

4.2 Relation Classification

After the argument component detection, the next step is to determine which relations hold between the different components (Figure 4.1). Valid **BI** tag sequences from the previous step are extracted, which are then considered to be the argumentative components of one RCT. Those sequences are phrases and do not necessarily correspond to full sentences. The list of components then serves as input for the relation classification. The relation classification task can be tackled with different approaches. The option applied in this line of experiments is to treat it as a sequence classification problem, where the sequence consists of a pair of two components, and the task is to learn the relation between them. Contrary to approaches which try to build up pre-defined structures, such as argumentation schemes, here, the linking of each component is done independently without the constraint of a final argument structure, see Chapter 2.2. For this purpose, self-attending transformers are employed, since these models are dominating the benchmarks for tasks which involve classifying the status between two sentences [59]. Treating it as a sequence classification problem opens up two options to model the problem: *(i)* jointly modelling the relations by classifying all possible argumentative component combinations or *(ii)* predicting possible link candidates for each entity and then classifying the relation only for plausible entity pairs. In the literature, both methods are represented. Therefore, both ways of solving the problem are evaluated. The here conducted experiments investigate various transformer architectures and compare them with state-of-the-art AM models, i.e., the Tree-LSTM based end-to-end system from Miwa and Bansal [106] as evaluated for AM by Eger et al. [36], and the multi-objective residual network of Galassi et al. [107].

Sequence Classification

For option *(i)*, a list of all the contained components is created for each abstract, respectively. From this list each component is exhaustively paired with all possible other components, resulting in $n-1$ component pairs per component, where n is the length of the list. As the model architecture, bi-directional transformers [59] are employed, which consists of an encoder and decoder which themselves consists of multi-head self-attention layer each followed by a fully-connected dense layer. Contrary to the sequence tagging transformer, where each token of the sequence has a representation which is fed into the classification layer, for sequence classification a pooled representation of the whole sequence is needed. This representation is passed into a linear layer with a softmax which decodes it into a distribution over the target classes. In the case of relation classification, the input does not consist of a single sentence, as it was the case for sequence tagging, but of the component pair separated by a special token. Similar to single sentence classification, the single pooled representation (of the component pair) is then passed to the classifying layer. Given that the *partial-attack* and *attack* labels are merged, because of their rare occurrences, this results in a three class classification problem (*Support*, *Attack* and *NoRelation*). In the following, this type of transformer is referred to as **SentClf**. Using this architecture one component can have relations with multiple other components, since each component combination is classified independently. There are various ways in constraining pre- or post-processing of the component pairs to eschew creating divergent argument structure. Treating it as a multiple choice problem is another way to implicitly limit the created structure to be convergent.

Multiple Choice

In a multiple choice setting (**MultiChoice**) the possible links are predicted taking the other combinations into account. This problem formulation is employed to address *(ii)*, i.e., first finding possible link candidates and subsequently classifying plausible combinations as *attack* or *support*.

In particular, each component (source) is given the list of all the other components as possible target relation candidates and the goal is to determine the most probable candidate as a target component from this list. This problem definition corresponds to the grounded common sense inference problem [108]. To model components which have no outgoing link to other components, the *noLink* option is added to the choice selection. As an encoder for phrase pairs, various BERT models, which are explained in the transformers section, are evaluated, just as for the SentClf task. With respect to the neural transformer architecture, a multiple choice setting means that each choice is represented by a vector $C_i \in \mathbb{R}^H$, where H is the hidden size of the output of an encoder. The trainable weight is a vector $V \in \mathbb{R}^H$ whose dot product with the choice vector C_i is the score of the choice. The probability distribution over

all possible choices is given by the softmax, where n is the number of choices:

$$P_i = \frac{e^{V \cdot C_i}}{\sum_{j=1}^{n} e^{V \cdot C_j}} \tag{4.3}$$

Since the softmax considers the score (dot product) of each choice, the order of the choices is ignored and does not bias the model. This step only provides a measure of how likely it is that the two components have a relation without specifying which type of relation it is. Subsequently, for each abstract, the component combination with the highest score of having a link between them is passed into a linear layer to determine which kind of relation is holding between the two components, i.e., *Attack* or *Support*. The MultiChoice model is trained jointly with two losses, i.e., one for the multiple choice task and one for for the relation classification task. To have a fair comparison with the SentClf, the predictions of the MultiChoice are translated into all possible combinations between components ($n^2 - n$ combinations for a list length of n), which is precisely the same what the SentClf is evaluated on. Concretely, this means that all component pairs, but the one with the highest score in the multiple choice step of the MultiChoice model which is passed to the relation classification step, are classified as having *noRelation*.

Linear Models

Furthermore, experiments with linear options for link prediction, such as matrix or tensor factorization were conducted. Those methods are widely used on graph data, e.g., knowledge graphs, to discover new links between existing nodes [109]. The matrix or tensor representation of the graph data is decomposed and a model specific scoring function, which assigns a score to each triple[4], is minimized, like a loss function in neural architectures. The goal was to combine those graph-based embeddings and enrich the nodes with linguistic features/embeddings to learn hybrid graph embeddings for relations and discover new links between arguments. The tested linear models are: TuckER [110], TransE [111] and ComplEX [109]. Unfortunately, those models did not learn a meaningful relation representation. This might be due to the relatively small graph data, which can be constructed from the AbstRCT corpus. In the literature, the smallest dataset these models have been experimented on has around 93k triples [112], whereas the AbstRCT dataset has less than 20k.

4.2.1 Experimental Setup

For fine-tuning the BERT model, the uncased base model with 12 transformer blocks, a hidden size of 768, 12 attention heads, a learning rate of 2e-5 with Adam optimizer was used. The pre-trained models were fine-tuned for 3 epochs. The same

[4] A triple consists of a subject (source node), a predicate (labeled edge between nodes) and an object (target node).

Method	Neoplasm	Glaucoma	Mixed
Tree-LSTM	.37	.44	.39
Residual network	.42	.38	.43
BERT MultiChoice	.58	.56	.55
BioBERT MultiChoice	.61	.58	.57
SciBERT MultiChoice	.63	.59	.60
BERT SentClf	.62	.53	.66
BioBERT SentClf	.64	.58	.61
SciBERT SentClf	**.68**	.62	**.69**
RoBERTa	.67	**.66**	.67

TABLE 4.5: Results of the relation classification task, given in macro F_1-score.

configuration was used for fine-tuning Sci- and BioBERT. Similar to the sequence tagging, for SciBERT, the uncased model with the SciBERT vocabulary was chosen, while BioBERT is a cased model. For BioBERT, version 1.1 is used. Additionally, the freshly released **RoBERTa** [67], another newer model, which outperforms BERT on the General Language Understanding Evaluation (GLUE) benchmark, was add to the selection of models. There, the BERT pre-training procedure is modified by exchanging static with dynamic masking, using larger byte-pair encoding and batches size, and increasing the size of the dataset. For RoBERTa, the number of epochs for fine-tuning was increased to 10, as it was done in the original paper. The best learning rate was 3e-5 on the SentClf task. The number of choices for training the multiple choice model was set to 6. Batch size was 8 with a maximum sequence length of 256 sub-word tokens per input example. Dataset splits were exactly the same as for the sequence tagging task, i.e., a neoplasm training set of 350 abstracts, a neoplasm development set with 50 abstracts, a neoplasm test set with 100 abstracts, a glaucoma test set with 100 abstracts and a mixed test set comprising 20 abstracts of neoplasm, glaucoma, hypertension, hepatitis and diabetes, respectively.

4.2.2 Results and Discussion

The results for relation classification are shown in Table 4.5. Results are given on all three test sets in macro multi-class F_1-score.

The Tree-LSTM based system performed the worst with a F_1-score of .37. This can be explained by the positional encoding in the persuasive essay dataset being more relevant than for clinical trials. There, components are likely to link to a neighboring component, whereas in the RCT dataset the position of a component only partially plays a role, and therefore the distance in the dependency tree is not a meaningful feature. Furthermore, the authors specify that their system does not scale with increasing text length [36]. Especially detailed reports of measurements can make RCT abstracts quite long, such that this system becomes not applicable for this type of data. The residual network [107] performed better with a F_1-score of .42. The main

problem here is that it learns a multi-objective for link prediction, relation classification and type classification for source and target component. Each task allocates capacities in the network. In the proposed AM pipeline the latter classification step is already covered by the sequence tagger and therefore unnecessarily repetitive at this step. Similar to sequence tagging, one can see a notable increase in performance when applying a BERT model. Comparing the specialized and general BERT model, the Bio- and SciBERT increase the performance by up to .06 F_1-score. Interestingly, RoBERTa delivers comparable results even though it is a model trained on general data. The speculations are that parts of the web crawl data which was used to train RoBERTa contain PubMed articles, since they are freely available on the web. Independently of that, RoBERTa shows more reliable results when looking at the performance on the out of domain test sets. While SciBERT, as the best performing system on the in-domain test set, drops .06 points on the glaucoma test set, RoBERTa stays almost the same and only drops from .67 to .66 F_1-score. Looking at the difference between the MultiChoice and SentClf architectures, the SentClf delivers better results, but the drawback is that this technique tends to link components to multiple components. Since most of the components in the AbstRCT corpus have only one outgoing edge, it creates a lot of false positives, i.e., links which do not exist. A problem with the MultiChoice is also the *noLink*. Since the input requires a sentence pair, but *noLink* means there is no second component, only the source component is fed into the classifier. It was meant to detect components, mostly claims, that are root nodes in the argument graph. Practically, the model could not learn a meaningful pattern to recognize those root nodes efficiently. While the AbstRCT dataset consists of only study abstracts for practical reasons, the pipeline can be applied on full text articles as well. Alas, a quantitative analysis on full articles cannot be provided due to missing annotated data. In preliminary experiments on full articles, a notable increase of false positives in the relation classification was observed, which is the expected consequence of an increased number of components. Furthermore, with the number of components rising in the double-digit range, the multiple choice architecture loses its predictive power. Further investigations to determine the exact limit of this architecture applied on full text articles is left to future work when annotated data is available.

Error Analysis Concerning link prediction, general components like *the difference was not statistically significant* are problematic, since it could be linked to most of the components/outcomes of the trial. Here, a positional distance encoding could be beneficial, since those components are usually connected to the previous component. In general, most of the errors in the MultiChoice architecture were made in the multiple choice part by predicting a wrong link and not at the stage of classifying the relation type. Interestingly, comparing the two domain adapted models, Bio- and SciBERT, there were no regular errors, which allows any conclusion about the advantages or disadvantages of one model.

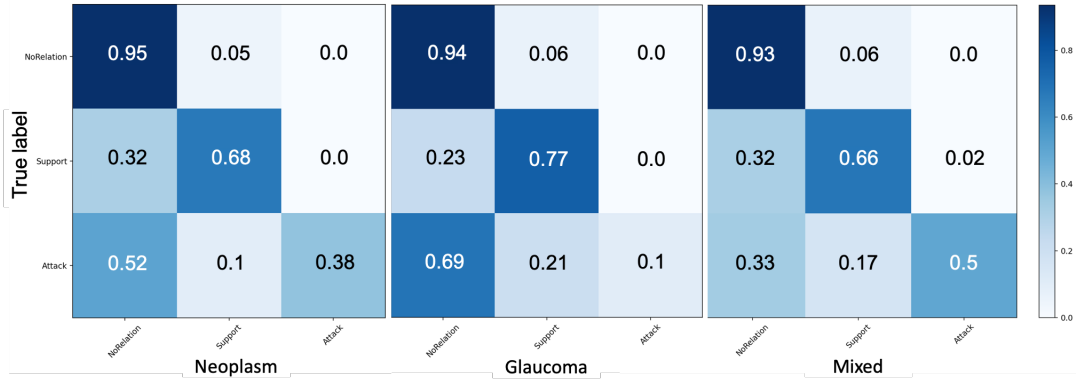

FIGURE 4.3: Confusion matrices of the predictions on the test set (neoplasm, glaucoma, mixed) of the relation classification task.

Looking at the confusion matrices, all tested SentClf models show a higher misclassification towards the *NoRelation* class. The confusion matrices of the SciBERT SentClf on all three test sets are shown exemplary in Figure 4.3. It can be further observed that the model could not learn a meaningful representation of the underrepresented *Attack* class. Most of the attack relations were not detected and classified as *NoRelation*. Similarly, the *Support* relation was mostly not confused as *Attack*, but as *NoRelation*. These false negative errors indicate that the model is overly focusing on the *NoRelation* class. This is in line with the observations from the MultiChoice approach, that the problem is in the multiple choice part, i.e., finding the right links between components.

Concerning the learned representation of the relation classes, both transformer approaches have in common the problem of dealing with negations and limitations or associating the polarity of a measurement and therefore confusing support and attack, which might indicate that the model learns rather linguistic patterns than a deeper understanding of the components and their relations. This would be in line with the examination of Niven and Kao [113], which applied BERT to the Argument Reasoning Comprehension Task (ARCT) [113] and found that the transformer is a strong learner for linguistic cues, but not for argument comprehension.

Example 4.2.1 [more research about the exact components of a VR intervention and choice of outcomes to measure effectiveness is required] $\xrightarrow[\text{Attack}]{\text{Support}}$ **[Conducting a pragmatic trial of effectiveness of a VR intervention among cancer survivors is both feasible and acceptable]**

Example 4.2.2 [this did not translate into improved progression-free survival (PFS) or overall survival] $\xrightarrow[\text{Support}]{\text{Attack}}$ **[The addition of gemcitabine to carboplatin plus paclitaxel increased treatment burden, reduced PFS time, and did not improve OS in patients with advanced epithelial ovarian cancer]**

Example 4.2.1 shows two claims with a limiting/attacking relation, which was wrongly classified as supporting. Concerning the polarity of an outcome, in Example 4.2.2, *not improving progression-free survival (PFS)* corresponds to a *reduced PFS time*, while for other factors reducing the value means it is beneficial and therefore improving some study parameter. This problem was also observed by Green [51], which found that the warrants for biomedical augmentation are often implicit. In the context of RCTs, for instance, the implicit warrant is that a certain value needs to be reduced to improve the overall result. Here, the inclusion of external expert knowledge is crucial to learn these fine nuances and compensate for the implicit warrant to correctly identify the relation. In this sense, the polarity of a measurement cannot be learned from textual features alone. Especially in the medical domain, where complex interrelationships are often implicitly presumed and therefore are impossible to capture with a model trained solely on character-based input. Phrases like *increased the blood pressure by X* or *showed no symptom of Y* can connote different messages depending on the context. Future work needs to consider this challenge of incorporating external expert knowledge. While I do not think this is a problem limited to a special domain, I consider it greatly important for understanding and representing medical text.

Chapter 5

Evidence Type Classification

This chapter introduces a subtask of argument component detection, i.e., evidence type classification. A further classification of the evidence is worthwhile, because in EBM the results of a clinical trial are rated based on various factors. Hence, to model this variety, the pieces of evidence in the AbstRCT dataset are assigned with specific evidence type labels, in particular, the more fine-grained label comparative, significance, side-effect and other. Various classification models including SVMs and NNs are evaluated on this task. This chapter comprises the work published at the Argument Mining workshop co-located with EMNLP-2018 [43].

The evidence from RCTs can be manifold and the ability to automatically extract the *arguments* proposed therein can be of valuable support for clinicians and practitioners in their daily evidence-based decision making activities. Given the peculiarity of the medical domain and the required level of detail, the standard approach to argument component detection in AM, as it was applied in the preceding chapter, is not fine-grained enough to fully support such activities. As a consequence, in my work the detected argument components are enriched with more information. In this chapter, a more fine-grained annotation scheme is proposed to distinguish different kinds of evidence in RCTs, so that fine-grained evidence-based decision making activities are supported. This is defined as a subtask of the argument component identification and called *evidence type classification*. The distinction among different kinds of evidence is crucial in evidence-based decision making as different kinds of evidence are associated to different weights in the reasoning process. For example, the recommendation based on aggregated evidence described by Hunter and Williams [8][1] builds upon preference settings, which are determined by the type of evidence. As it is further discussed in Chapter 9.1.2, detecting and evaluating comparisons are targets of high interest and common precursor tasks in the domain. To extract this information, which is contained in but not expressed with the coarse argument component labels, I propose four new classes of evidence for RCT (i.e.,

[1] This approach is described in detail in Chapter 9.1.1.

comparative, *significance*, *side-effect*, and *other*). Previous work on evidence classification [97] tackled the problem on Wikipedia based data, dividing the evidence into *study, anecdotal and expert* evidence. While this taxonomy could be applied on a higher level in the decision making process if other evidence is considered, such as personal experience of the practitioner, it is not applicable for the here presented type of data or clinical trials in general. Additionally to the presented work and responding to the feedback from medical experts, the argument components are further aggregated with the Effect-on-Outcome, which is described in Chapter 6.

To address *evidence type classification*, a supervised approach is proposed and tested on a set of RCT abstracts on different medical topics. Section 5.1 describes the annotation scheme and the rationale behind the choice of labels. Subsequently, the proposed methods are presented in Section 5.2, and results are discussed in Section 5.3.

5.1 Annotation Scheme

This work was conducted before the data collection and extension of the AbstRCT dataset to version 2 (see Chapter 3). Thus, the fine-grained annotations and the experiments based on it comprise only the abstracts from the first version of the AbstRCT dataset containing 169 abstracts in total.

As a quick reminder, an *evidence* in a RCT is an observation or measurement in the study, which supports or attacks another argument component, usually a *claim*. They are observed facts, and therefore credible without further justifications, since this is the ground truth the argumentation is based on. The coarse evidence label comprises indiscriminately observations like side effects and the measured outcome of the intervention and control arm. Example 5.1.1, *evidence* are in italic, underlined and surrounded by square brackets with subscripts, while claims are in bold.

Example 5.1.1 *To compare the intraocular pressure-lowering effect of latanoprost with that of dorzolamide when added to timolol. [...] [The diurnal intraocular pressure reduction was significant in both groups ($P < 0.001$)]$_1$. [The mean intraocular pressure reduction from baseline was 32% for the latanoprost plus timolol group and 20% for the dorzolamide plus timolol group]$_2$. [The least square estimate of the mean diurnal intraocular pressure reduction after 3 months was -7.06 mm Hg in the latanoprost plus timolol group and -4.44 mm Hg in the dorzolamide plus timolol group ($P < 0.001$)]$_3$. Drugs administered in both treatment groups were well tolerated. This study clearly showed that* ***[the additive diurnal intraocular pressure-lowering effect of latanoprost is superior to that of dorzolamide in patients treated with timolol]$_1$.***

Different reports of the experimental outcomes as evidence can be observed in this example. Those can be results without concrete measurement values (see Evidence 1), or exact measured values (see Evidence 2 and 3). Different measures are annotated as multiple evidence. The reporting of side effects and negative observations

are also considered as evidence. To further distinguish evidence into finer classes, one first has to reflect which purpose the labels should fulfill. Traditionally evidence-based medicine focuses mainly on the study design and risk of bias, when it comes to determining the quality of the evidence. As stated by Bellomo and Bagshaw [114], there are also other aspects of the trial quality, which impinge upon the truthfulness of the findings and should be considered in the critical appraisal of evidence and interpretation of the results. As a step forward, the dataset annotation was extended, specifying four classes of *evidence*, which are the most prominent in the AbstRCT data and assist in assessing these complex quality dimensions, like reproducibility, generalizability or the estimate of effect:

- ***comparative*:** when there is some kind of comparison between the control and intervention arms (Table 5.1, Example 2), supporting the search for similarities in outcomes of different studies, which is an important measure for the reproducibility. Due to the comparative nature of RCTs, this label is more frequent than the others.
- ***significance*:** for any sentence stating that the results are statistically significant (Table 5.1, Example 3). Many comparative sentences also contain statistical information. However, this class can be seen more as a measure for the strength of beneficial or potentially harmful outcomes.
- ***side-effect*:** captures all evidence reporting any side-effect or adverse drug effect to see if potential harms outweigh the benefits of an intervention (Table 5.1, Example 4).
- ***other*:** all the evidence that do not fall under the other categories, like non-comparative observations, risk factors or limitations of the study (too rare occurrences to form new classes). Especially the latter can be relevant for the generalizability of the outcome of a study (Table 5.1, Example 5).

Comparative structures are important means in scientific communication and an essential part of clinical trials. Thus, previous work in the domain [115, 116] investigated ways of automatically detecting and evaluating these structures (for more details, see Chapter 9.1.2). Since all comparisons related to the outcomes of a study are covered within the argumentative components, it naturally made sense to create this more fine-grained label to highlight these *comparative* structures. In combination with the Effect-on-Outcome and PICO element detection, comparisons are found, marked, evaluated and put in a structured format, to ease querying and analysing such data. Concerning the *significance* class, this label is related to the statistical significance and generalizability of a trial. Only statistical significant outcomes should be considered to draw conclusions from a trial. Moreover, this class comprises statements that an observation did not reach statistical significance. Even more important

1.	**Claim:** Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response.
2.	**Comparative:** The overall success rates were 87% for the 350-mm2 group and 70% for the 500-mm2 group (P=0.05).
3.	**Significance:** All regimens produced clinically relevant and statistically significant ($P < .05$) intraocular pressure reductions from baseline.
4.	**Side-effect:** Allergy was seen in 9 % of subjects treated with brimonidine.
5.	**Other:** Risk of all three outcomes was higher for participants with chronic kidney disease or frailty.

TABLE 5.1: Sample of the positive classes represented in the corpus for evidence type classification (*Claim*, *Comparative*, *Significance*, *Side-effect*, *other*).

than statistical significance, which states that the observed effect on an outcome was not by chance, is clinical significance or estimate of effect. Clinical significance sets the study results into perspective. Is the expected benefit of an intervention worth the risks coming along with this intervention? Are the costs and the required effort proportional to what is gained? These estimates can be determined on a general level, e.g., if a treatment should be recommended in clinical guidelines, but also on an individual level for each patient case. Evaluating the trade-off is a highly complex task, even for experienced practitioners. Many factors and circumstances come into playing a role. Modelling and automatically evaluating this is even harder and requires vast domain knowledge. Thus, the goal of the thesis is not to make the decision for the practitioners, but to provide as much adjuvant information as possible. With the previous two labels targeting other desiderata, the *side-effect* label should contribute to the question of clinical significance. This label is a means to highlight potential risks, which is important for the evaluation of a treatment. Even statements about the non-existence of adversarial effects are valuable sources of information and thus, are contained within this label. It would have been preferable to have more classes like this targeting the estimate of effect. There are also reports about risk factors and limitations of a study in the data. Unfortunately, the amount of these is too small to justify separate classes. In a taxonomy, these labels should be distinctively included, however, given that the objective was also to practically apply this annotation scheme on the data and build a functioning classifier, this forced the decision of merging them into the *other* class.

Dataset	*Topic*	*#abstract*	*#comp.*	*#sign.*	*#side-eff.*	*#other*
Training set	glaucoma	79	151	83	65	10
Test set	glaucoma, diabetes, hepatitis, hypertension	90	160	98	79	33

TABLE 5.2: Statistics on the evidence type dataset showing the class distributions.

As previously stated, the annotations were executed on the first version of the AbstRCT corpus. Table 5.2 shows the statistics of the obtained dataset. With 49% and 43% respectively, the *comparative* is evidently the dominant class in the dataset.

While the *significance* and *side-effect* labels are within a comparable range to each other, the *other* class is underrepresented from a machine learning point of view, which is expected given the definition of this class. As for the dataset creation, three raters have annotated the data after a training phase. In line with previous and subsequent annotation phases, the inter-annotator agreement has been calculated with three annotators on a previously unseen subset of the data. For the evidence type annotation, this subset consisted of 10 abstracts comprising 47 evidence, resulting in a Fleiss' kappa of 0.88, attesting the reliability of the guidelines and the obtained dataset.

5.2 Experimental Setup

In work contemporary to the approach I present in this chapter, I addressed the argument component detection as a supervised text classification problem [79]: given a collection of sentences, each labeled with the presence/absence of an argument component, the goal is to learn a discrimination function $f : X \rightarrow Y$ to infer the label from the input text. For the first step of the AM pipeline, i.e., the argument component classification, I decided to rely on an existing system and to tailor it to cope with the clinical data scenario. More precisely, an existing system, i.e., MARGOT [95], is re-trained to detect evidence and claims from clinical data (for more details, see Chapter 4.1.1). To this end, SubSet Tree Kernels (SSTK) [98] were used, which offer a reasonable compromise between expressiveness and efficiency [95]. In SSTK, a fragment can be any sub-tree of the original tree, which terminates either at the level of pre-terminal symbols or at the leaves. Data was pre-processed (tokenisation and stemming), and the constituency parse tree for each sentence was computed. Furthermore, the bag-of-words features with tf-idf values were also computed. All the pre-processing steps were performed with Stanford CoreNLP (version 3.5.0). The experiments were conducted with different classifiers and feature combinations. Two datasets were prepared to train two binary classifiers for each approach: one for claim detection, and one for evidence detection. Both training sets only differ in the labels, which were assigned to each sentence. 5-fold cross validation was performed optimizing for the F_1-score. The model was evaluated on the test set in Table 5.2 obtaining 0.80 and 0.65 F_1-score for evidence and claim detection respectively.

As a step forward – after the distinction between argumentative (claims and evidence) and non-argumentative sentences – I addressed the task of distinguishing the different types of evidence. It was cast as a multi-class classification problem. For that SVMs[2] with a linear kernel were used. Since SVMs were designed for binary classification they do not natively support multi-class classifications. There are different strategies to transform the multi-class into a binary classification problem: *(i)* ONEVSREST, and *(ii)* ONEVSONE. The first strategy trains one classifier for each class, where the negative examples are all the other classes combined, outputting

[2]scikit-learn, version 0.19.1

a confidence score later used for the final decision. The second one trains a classifier for each class pair and only uses the correspondent subset of the data for that. Both strategies were evaluated in the experiments. As features for the SVM, lexical ones, like tf-idf values for bag-of-words, n-grams and the MedDRA[3] dictionary for adverse drug effects were selected. The models were compared against a random baseline, based on the class distribution in the training set and a majority vote classifier, which always assigns the label of the class with the highest contingent in the training set. In later experiments, these models were compared against neural recurrent architectures for sentence classification. Here, the input is encoded via word embeddings and subsequently passed through a GRU. The full text representation from the GRU is passed through a final linear layer for classification. Different encoding methods for the input were evaluated, e.g., static word embeddings considering full words (GloVe) and sub-words (FastText), and contextualized embeddings (ELMo). As for the argument component classification, the models were evaluated on different test sets with respect to the weighted average F_1-score for multi-class classification. Here, the score is weighted by the support, which is the number of true instances for each label. The first dataset consists only of the glaucoma data, and the second one comprises all the other maladies in the dataset as well (see Table 5.2).

5.3 Results and Discussion

The experiments were conducted in two application scenarios. In the first scenario, the evidence type classifier was tested on the gold standard annotations of evidence in the RCT dataset. This excludes all claims and non-argumentative sentences from the experiments. In the second scenario, the whole pipeline is tested: the evidence type classifier is run on the output of the aforementioned argument component classifier. For the neural architectures, the component classifier is integrated directly. This translates into a six class classification problem, where contrary to the gold standard approach, the *claim* and *non-argumentative* classes are considered. Similar to the antecedent argument component classification and established practice for SVMs, the best feature combinations were selected. In both scenarios, the best feature combination was a mix of bag-of-words and bigrams. The dictionary of adverse drug effects did not increase the performance. Together with the fact that the data contains just a small group of reoccurring side-effects, this suggests that the expected discriminative information from the dictionary is captured within the uni- and bigram features. This might change for bigger datasets with a broader range of adverse effects.

Results For the evidence type classifier on gold standard annotations, the obtained results regarding the different multi-class strategies did not differ significantly, as

[3] https://www.meddra.org/

can be observed in Table 5.3. The results of the random baseline, the SVM with the best feature combination, and neural models are reported in Table 5.4.

Dataset	*Strategy*	*glaucoma*	*combined.*
Gold standard	ONEVSREST	.80	.73
	ONEVSONE	.79	.74
whole pipeline	ONEVSREST	.71	.65
	ONEVSONE	.71	.66

TABLE 5.3: Results of the two multi-class strategies for the evidence type classifier (SVM with best features) in weighted f_1-score.

For the classification on the evidence gold standard, the SVM performed best achieving a F_1-score of .80 and .74, respectively for the glaucoma and combined test set. Reviewing the best n-grams, they contain very specific medical terminology, explaining the performance difference between the two test sets. As a possible future extension, another pre-processing step with better abstraction capability, e.g., substituting concrete medical related terms with more general tags, could provide benefits for the trained model on the out-of-domain task. Interestingly, the neural models do not perform as well as the SVM. This might be due to the lack of training data. Since NNs learn patterns for all classes jointly, they require more training data than a SVM, which casts the problem into multiple binary problems and therefore has a better class specific discrimination capability given the bigger size of negative samples for each binary task. This changes for the classification over all argumentative labels. Here, the F_1-score of the SVM is .71 for the glaucoma and .66 for the combined test set, which is lower than the ones from the neural networks. Adding the two classes tripled the size of the training data from 309 to 945 examples, so that the NNs could start unfolding their power. For the SVM, as expected, the errors of the argument component classifier have an impact on the performances of the second step lowering the results with respect to the performance on gold standard, but that setup corresponds to a more realistic scenario.

Dataset	*Method*	*glaucoma*	*combined.*
Gold standard	RANDOM	.33	.32
	MAJORITY	.27	.26
	N-GRAMS	**.80**	**.74**
	GLOVE	.60	.41
	FASTTEXT	.75	.60
	ELMO	.73	.57
whole pipeline	RANDOM	.38	.38
	MAJORITY	.38	.39
	N-GRAMS	.71	.66
	GLOVE	.73	.65
	FASTTEXT	.78	.70
	ELMO	**.80**	**.71**

TABLE 5.4: Results of the argument component detection on AbstRCT v1 (weighted average F_1-score).

As for the comparison of the various word embeddings, GloVe embeddings resulted in the lowest F_1-score with .73, which is only marginally higher than the n-gram based SVM. In line with the observations in Chapter 4.1.2, the sub-word based fastText and the contextualized ELMo encode the input text with a comparable quality, where ELMo is only marginally better. Concerning the generalizabilty and transfer capabilities for the out of domain test set, a notable drop in performance is registered for all word embeddings. While also the SVM shows the common decreased in performance for out of domain data, the magnitude of the dip is not as great as for the neural network. This could be explained with the hypothesis that a certain amount of examples are clearly identifiable also for the out of domain data. But given that the performance dip is consistent over all embedding types, even for the higher performing ones, the more likely explanation is that for low data scenarios, SVMs are still a competitive solution.

Error Analysis As shown in Figure 5.1, *side-effects* were often confused as *comparative*. Certain types of *side-effect* comprise comparisons of side-effects between the two groups. This includes statements of the non-existence of adverse reaction. The structure and wording of those sentences are very similar to correct *comparative* examples and only differ in the comparison criteria (side-effect vs. other measurement), see Examples 5.3.1 and 5.3.2 as instances of this misclassification. Furthermore, *comparative* and *significance* labels were often confused. As explained earlier, comparisons can also state information about the statistical significance and could therefore belong to both classes, see Example 5.3.3. A possible solution to overcome this problem in the future could be a multi-label approach to assign more than one attribute to a piece of evidence.

Example 5.3.1 (**predicted label:** Comparative; **correct label:** Side-effect)
Headache, fatigue, and drowsiness were similar in the 2 groups.

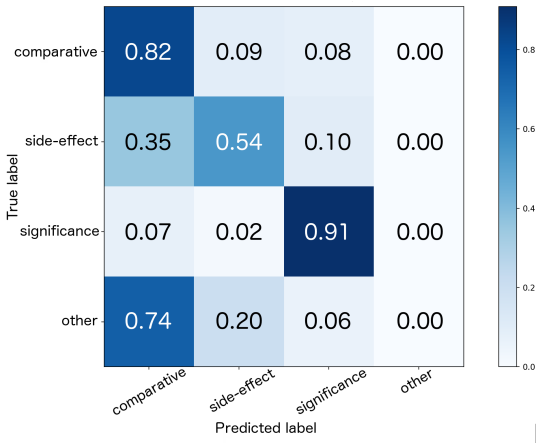

FIGURE 5.1: Normalized confusion matrix of the predictions of the SVM for evidence type classification on the combined test set.

Example 5.3.2 (**predicted label:** Comparative; **correct label:** Side-effect)
The number of adverse events did not differ between treatment groups, with a mean (SD) of 0.21 (0.65) for the standard group and 0.32 (0.75) for the intensive group (P=0.44).

Example 5.3.3 (**predicted label:** Significance; **correct label:** Comparative)
The clinical success rate was 86.2% in the brimonidine group and 81.8% in the timolol group, making no statistically significant difference between them (p=0.817).

To summarize, the proposed fine-grained labels are an important step towards structuring the extracted evidence. For instance, they can be used to associate different weights to the single pieces of evidence in the reasoning process of an argument-based decision system. Further, they carry valuable information for clinicians who want to get an overview of a clinical trial. They make the provided information more detailed and allow filtering for certain categories, e.g. side-effects, which becomes even more handy with the subsequently described analysis of outcomes.

Chapter 6

Effect-on-Outcome Analysis

This chapter introduces the analysis of the results of a clinical trial. Specifically, the analysis of the effect of an intervention on the observed outcome parameters. In the following, this is called the Effect-on-Outcome extension of the Argument Mining pipeline. To this end, outcomes mentioned in the argumentative components are detected and their effects are classified to assess if an intervention has Improved, Increased or Decreased the outcome, or that there was NoDifference, or NoOccurrence of the outcome.

So far, the Argument Mining pipeline spans the detection of components and their boundaries, their classification in (fine-grained) classes, and the argument structure prediction. This already gives a good overview of the clinical trial. However, crucial information about the results of a study are still only available in a human-readable format. Thus, in this chapter, I propose a way of analysing the effects of interventions encoded in the argument components in a way that it can be easily translated into machine-processable data. As described in Chapter 2.1, in EBM PICO elements play an important role. Hence, integrating PICO elements in the argumentative structure seemed the next logical step in expanding the pipeline. The first step towards this was taken in the context of the ACTA demo system, which is described in detail in Chapter 8.1. With the EBM-NLP dataset [6] being freshly released, there was a sizable resource available, which could be used to train models for PICO element extraction. Technically, the PICO element extraction can be formulated as a sequence tagging problem, similar to the argument component detection. Already having a sequence tagging architecture at hand and with the new EBM-NLP dataset available, a PICO extraction model was trained, achieving a F_1-score of 0.73 on the EBM-NLP test set for coarse labels. The model was trained jointly on the coarse label version of the dataset, providing a tool to predict participant, intervention and outcome candidates in an RCT. Intervention and the comparison intervention[1] are not considered as two separated labels, since they comprise the same vocabulary and the right label is based on the function in the trial, which cannot always be inferred

[1] As a quick reminder, in clinical trials researchers aim at comparing a (new) intervention (the *I* in PICO) against established (comparison) interventions or placebos (the *C* in PICO).

from single isolated sentences. This would require a more sophisticated analysis of the whole abstract. Nevertheless, the additional information provided in form of PI(C)O elements resulted in positive feedback from medical experts, which encouraged me to deeper entwine PICO information with the argument graph.

With ACTA, the PICO detection was not specifically targeted at argumentative components. During the detection, the abstract as a whole is annotated with the elements independently of the found arguments. This was done mainly because information about the population is usually not available in argumentative components, i.e., approximately 1-2% contain information about the trial population. Thus, processing only the argumentative components in the PICO detection would have basically circumvented the extraction of information about the trial participants, which is not desirable.

With a first approach to PICO detection, the next step is to analyse the effect an intervention has on an outcome. As shown in Chapter 9.1.2, finding comparative sentences was motivated by the fact that the aspect of comparing interventions with respect to a certain outcome is an imperative part of EBM practice. Hence, an automatic analysis of the Effect-on-Outcome would add to the versatility of the application of the AM pipeline, not only as additional information for the practitioner, but also to create a richer structured input for argumentation-based assistance systems, as described in Chapter 9.1.1. For these reasons, as an extension of the classic Argument Mining pipeline, an automatic outcome analysis is integrated to enrich the arguments with valuable medical information and leverage this way the advantages of both domains.

The proposed method how the outcome analysis is addressed is introduced in Section 6.1 and the specifications of the experimental setup are stated in Section 6.2. Subsequently, in Section 6.3 the results are presented and observed problems discussed.

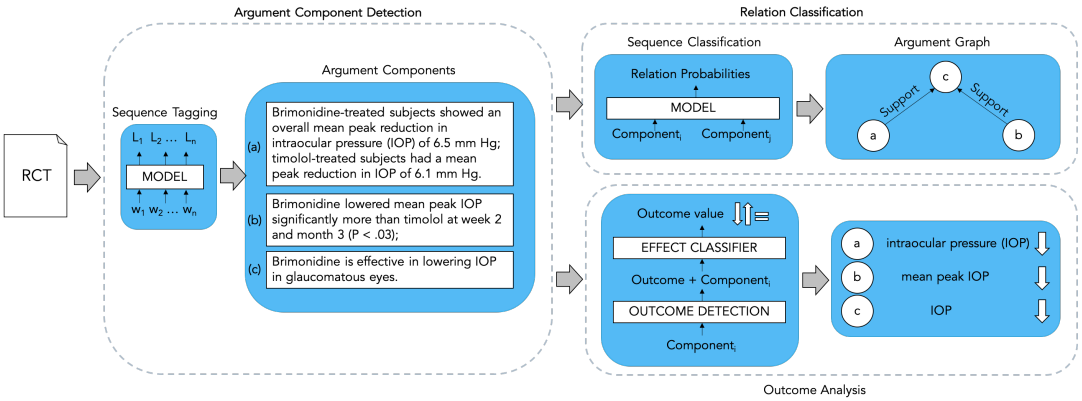

FIGURE 6.1: Illustration of the full Argument Mining pipeline with the outcome analysis extension.

6.1 Outcome Analysis Pipeline

Based on the annotation guidelines for the AbstRCT dataset, argumentative components should contain all mentions of Effect-on-Outcome, which are either in a comparative form or stated with respect to only one of the interventions. For this reason, the outcome analysis extension of the pipeline processes only phrases which were classified as argumentative components in the first step of the AM pipeline, and not every sentence of the abstract. In general, the outcome analysis is a pipeline itself with two major parts. First, an outcome detection, which finds and extracts the outcomes of an argumentative component, and second, an effect classifier, which predicts which consequence was seen for each outcome after an intervention. The role in the overall AM pipeline and the two parts of the outcome analysis are illustrated in Figure 6.1.

Similar to the argument component and PICO element detection, the outcome detection is treated as a sequence tagging task with the BIO-tagging scheme, resulting in a three class classification problem (*B-Outcome*, *I-Outcome* and *NoOutcome*). In accordance with previous experiments, the same transformer architecture for sequence tagging is employed with various alternatives for the pre-trained weights. After the classification step in this part of the pipeline, valid **BI**-sequences are extracted from the prediction results, which are considered to be the outcomes reported in a component. Each outcome is paired with the component it occurred in and serves as input for the effect classifier. Sentences with multiple detected outcomes generated multiple inputs, one for each detected outcome. Given this bipartite input, the problem is similar to the aforementioned relation classification, namely to sequence classification. Thus, the effect classifier follows the SentClf approach from Chapter 4.2. Differently from the three class relation classification, in this case, it is a five class (i.e., *Improved, Increased, Decreased, NoDifference, NoOccurrence*) classification problem.

6.2 Experimental Setup

Experiments are conducted with the same pre-trained transformer model types as for relation classification, i.e., **BERT**$_{base}$, **BioBERT** and **SciBERT** (cased and uncased), with the exception of RoBERTa. For both parts of the pipeline, i.e., the outcome detection and effect classifier, the same type of transformer is employed. As for the sequence tagging architecture the LSTM combination with a CRF was chosen for the experiments, because the difference between the LSTM and GRU approaches were only marginal for the argument component detection. The outcome pipeline implementation was done with the same Python, PyTorch and transformer versions as the previous experiments (see Chapter 4.1.2 and 4.2). Both transformer models of the pipeline are of the same type and initialized with the same pre-trained weights. The Effect-on-Outcome annotations are converted into two datasets, one for each part of

the pipeline. The first one in a CoNLL format for token-wise labels, and the second one in csv format, where each outcome-component pair is listed. This results in multiple entries, if a component contains more than one outcome. The fine-tuning of the models is done separately, each task on its own dataset version. The learning rate was set to 2e-5 with Adam optimizer and the models were fine-tuned over 3 epochs with a batch size of 32 and a maximal sequence length of 128 tokens. Token-wise evaluation is done on the full pipeline output, which is reconverted to CoNLL format to compare against the gold labels, taking the propagated error from the first pipeline part into account. The annotated dataset was split into a train and test set (80% and 20%, respectively) respecting the class distribution of the overall dataset in both subsets. Given the size of the dataset and that fact that the annotations are imbalanced with respect to certain classes (see Section 3.2), it is not feasible to maintain three test sets and ensure at the same time that they have the same size, as done for experiments on the AM pipeline (see Section 4.1.2), i.e., 100 abstracts each. Whilst it would be indeed interesting to see the effects the comparison of three different test sets offer, test sets with different sizes do not allow for a fair comparison.

6.3 Results and Discussion

The results for the outcome analysis pipeline are shown in Table 6.1. Results are given on the test set in macro multi-class F1-score and as a binary F_1-score for each of the five classes separately.

Model	F_1	Improved	Increased	Decreased	NoDiff	NoOcc
BERT (cased)	.62	.69	.65	.66	.75	.00
BERT (uncased)	.72	.72	.70	.72	.72	.50
BioBERT	.75	.74	.74	.77	.76	.54
SciBERT (cased)	.75	.71	.71	.73	.71	**.65**
SciBERT (uncased)	**.80**	**.81**	**.75**	**.81**	**.85**	.59

TABLE 6.1: Results for the outcome analysis pipeline, given in overall macro F_1 and label-wise binary F_1-score.

The baseline BERT$_{base}$ models perform the worst, with an interesting and unexpectedly worse performance of the cased model, in contrast with prior observations that the cased model performs better for case sensitive tasks [59], like NER. Apparently, the BERT$_{base}$ cased model is not capable of fully learning a representation of all classes, since the *NoOccurrence* class was not predicted a single time, resulting in an F_1-score of 0. Indeed, this class is underrepresented with respect to the other classes, but still, the other models were able to learn some patterns for it.

Similarly to the relation classification results, one can observe an increase in performance on the specialized Bio- and SciBERT models compared to the general BERT

model. In a direct comparison of the cased versions[2] of these two specialised models, the overall result is the same with .75 F_1-score. In the binary evaluation, BioBERT is slightly better with the exception of the *NoOccurrence* class. The SciBERT cased model performs the best with .65 F_1-score. The motivation behind the usage of cased models was to deal with outcome abbreviations, which are usually uppercase letters and relatively common in the data (45% of the argumentative components contain abbreviations). However, from the results no definite positive effect can be observed for cased models. In fact, these models seem to be unstable for underrepresented classes. Overall, SciBERT uncased is the best performing model with .80 macro F_1-score. It also outperforms the rest of the approaches in every F_1-score measured except for the *NoOccurrence* category, where the cased version has higher score. This category, in particular, suffers from sensitivity to class imbalance given that only 2% of the annotated data is labeled as such. For the other classes, the binary F_1-scores are in a comparable range to each other, where the most prominent class in the annotated data, i.e., *NoDifference* with 27%, has consistently the highest or second highest score. Besides the *noOccurrence* class, the *Increased* class has always the second lowest scores. Even for the best performing model, the difference compared to the worse performing models is not as massive as for the other classes. Notable in the confusion matrix, visualized in Figure 6.2, the classifier tends to wrongly predict it as *Improved*, which is a closely related class.

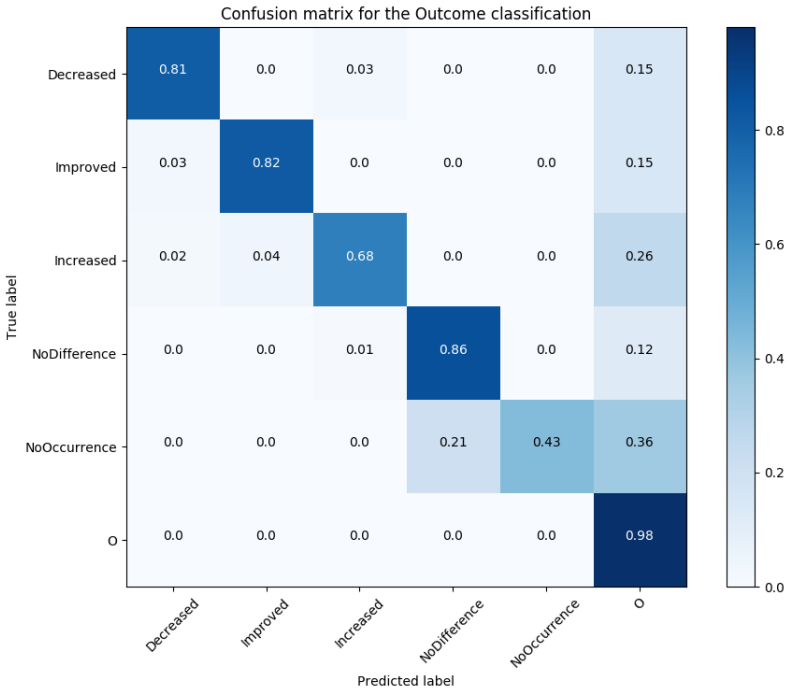

FIGURE 6.2: Confusion matrix of the predictions on the test set of the outcome classification.

[2]BioBERT is a cased model.

Error Analysis With respect to the source of error in the pipeline, the two pipeline parts cause different observable errors in the overall output. Being a binary classifier, the first part, the outcome detection, is the only part which predicts the negative class label (referred to as *O* in the confusion matrix). The second part, the effect classifier, assigns effect class labels (e.g., *Increased/Decreased*) to outcomes, which were found by the outcome detection module. Consequently, the impact of the propagated error from the first part of the pipeline can be observed in the confusion matrix in Figure 6.2. Effect classes are mostly not misclassified as other effect classes, but as the negative class *O*. This is reflected in a stronger coloration in the horizontal direction for the predicted *O* label in the confusion matrix. Since the only part in the pipeline which is responsible for the negative *O* label is the outcome detection, this means that the error occurred in the first part of the pipeline. Accordingly, confusion of effect class labels are errors of the second part of the pipeline, i.e., the effect classifier.

One of the most common mistakes of the models is the incomplete detection of outcomes. In many cases, the outcome to classify includes other words that complement it. For example, the outcome to detect in the sentence in Example 6.3.1 is *levels of VEGF*, while the model only catches *VEGF*.

Example 6.3.1 *The* levels of VEGF (PredictedSpan / CorrectSpan) *were significantly lower.*

Another common mistake occurs with the sentences that compare different results for the same outcome, the model does not have a clear reference point to perform a correct prediction. This can be seen in Example 6.3.2.

Example 6.3.2 *In* G1 / G2, *respectively, improvement/deterioration of* QoL (Decreased / Improved) *correlated with better or poorer* intake (Decreased / Improved) *or* nutritional status (Decreased / Improved).

If the reference point is the group study *G1*, the model should classify *QoL* as *Improved*, while should be *Decreased* for *G2*. The same applies to *intake* and *nutritional status*. In reality, the model confuses the reference points and predicts QoL as *Improved* and the rest as *Decreased*. For future work, an approach where the outcome is linked to the reference point by the model could be worth exploring.

Furthermore, it is found that the model is effectively tagging outcomes in such a way that the predicted labels are different from the true labels, but can be considered as correct nonetheless.

Example 6.3.3 Excess limb size (Decreased / Decreased) (circumference and water displacement (Decreased / NoOutcome)) *and* excess water composition (Decreased / Decreased) *were reduced significantly.*

The sentence in Example 6.3.3 has the outcomes *Excess limb size* and *excess water composition* as true outcomes, both labeled as *Decreased*. The model detects and classifies those outcomes correctly, but also adds the outcomes *circumferences* and *water displacement*, predicting the label *Decreased* which could be plausible labels, but were not annotated as such, since coordinations should be broken down in the smallest

possible units. However, in this case, they are not self-sustained on their own. *Circumferences* alone misses additional information, namely that it is the circumferences of limbs. An opposite case can be observed for modifiers, like *serious* or *severe*, which are not included in the outcome prediction, although they carry essential information.

To summarize this chapter, an extension of the Argument Mining pipeline was presented, i.e., the Effect-on-Outcome module. This module analyses the effect an intervention has on an outcome. To this end, in a first step, outcomes are detected in the argument components and subsequently examined in the effect classifier. For both parts of the pipeline, various transformer models were compared against each other. While in the preceding sections and chapters various errors of these models are discussed, the reasonable suspicion arose, particularly in Chapter 4.2, that the transformer models do not learn a deep understanding of the underlying semantics, especially for argument components and their relations. To investigate this further, I decided to explore possible manipulations of the input data with respect to changes in the prediction behaviour of the model. These results are presented in the following chapter.

Chapter 7

Robustness and Weaknesses of Transformer Models

As shown in previous chapters, whilst the employed LM based transformer models are pushing the state-of-the-art results, the question arose what exactly they are learning. To investigate the capability of those models to cope with variational input, this chapter introduces different ways of creating linguistically simple perturbations ranging from punctuation deletion to various word-based transformations. Their impact on the robustness of current state-of-the-art Language Model based argument classification models is evaluated, with respect to both in-domain and cross-topic performance. The quality of the generated perturbations is assessed in a user study and the effect of adversarial training for argument classification is empirically evaluated. Subsequently, other known weaknesses of LM based transformer models are highlighted. This chapter comprises the work published at the International Conference on Computational Models of Argument (COMMA-2020) [117] and the European Conference on Artificial Intelligence (ECAI-2020) [118].

In the last years, several empirical approaches have been proposed to tackle Argument Mining tasks, e.g., argument classification, relation prediction, argument synthesis. These approaches, as those presented in this thesis, rely more and more on Language Models (e.g., BERT) to boost their performance. However, these Language Models require a lot of training data, and size is often a drawback of the available Argument Mining data sets. The goal of the experiments presented in this chapter is to assess the *robustness* of these Language Models for low-resource tasks. One of these is the AM subtask of topic-dependent argument classification, where the goal is to find relevant arguments for a given topic or claim from heterogeneous sources. This task is currently addressed by employing state-of-the-art deep learning methods, that recently benefit from pre-trained Language Models like BERT [59]. As described in Chapter 2.3, the idea underlying LM pre-training is to learn a task-independent understanding of natural language in an unsupervised fashion, from

vast amounts of unlabeled text. After learning this general knowledge about a language, the model is then fine-tuned on tasks where the amount of available annotated data is significantly smaller. While the datasets used for the GLUE benchmark comprise still a decent amount of examples, AM datasets are considerably smaller in size. For instance, one of the biggest currently available AM datasets, i.e., the *UKP Sentential Argument Mining Corpus* [119], contains a little bit more than 25k examples, whereas the Stanford Natural Language Inference (SNLI) dataset [120] comprises over 570k examples, which is more than 20 times the size of the AM dataset. However, AM is a very context-dependent task and requires deep Natural Language Understanding with respect to the component detection and even more for the structure prediction. For the latter, the model does not only need to learn how argumentative statements look like, but also to which concepts and circumstances they exactly refer to. Moreover, the model needs to discover and learn the connection between two components to comprehend their interdependent meaning and infer their relationship to each other. In this context, the definition of Argument Mining algorithms, as targeted in the research questions in Chapter 1, extends to the analysis of how well the pre-trained NLU capabilities of the proposed model scale for fine-tuning on tasks with fewer resources available, such as argument classification. To this end, this chapter examines the vulnerability of argument classification models to adversarial attacks and adversarial training as a way of improving the robustness of a model. To address these issues, the efficiency of simple linguistic attacks against topic-dependent argument classification models based on LM pre-training are evaluated. Whilst this task is not explicitly addressed in the aforementioned AM pipeline for clinical trials, it could be used to formalize inter-trial debates about specific treatments. The argument classification subtask was chosen, because the topic domain and data structure of the AbstRCT dataset pose challenges, which make the evaluation of robustness on this dataset impractical. For example, linguistic changes in a sentence, which seem semantic preserving to the medical layman, can indeed impact and alter the pathological or therapeutic meaning. Moreover, adding perturbations to the sequence tagging input is hard. Perturbations are supposed to be minor change which preserve semantics, e.g., adding or replacing one word, which can cause a change in the prediction. Changing one single label in a sequence of labels, where the prediction of the label sequence is dependent on all states of the sequence (structured prediction), might not be a fruitful undertaking. Again, most of the proposed perturbation techniques are not easily applicable to the medical domain, as these changes could cause the (medical) meaning of a sentence to be altered, which perturbations should not do. Consequently, the verification of the preservation of semantics in this case would be cumbersome, because it would require trained medical experts to evaluate. To eschew this risk of changing semantics, I decided to evaluate the model on a task, where a quality control of the generated perturbations is feasible. For these reasons, the closest AM task, where BERT was state-of-the-art, was chosen, which is topic-dependent argument classification. Here, similarly to the

relation classification, the problem is a sequence classification task with the goal to determine the relation (argument for or against) between two input phrases. Thus, the obtained insights can give a general idea about the robustness of the underlying model to simple linguistic changes. Moreover, the human evaluation of the preservation of semantics can be done by non-medical experts, since the used datasets are about more general issues, commonly known or at least comprehensible for the reader.

Formally, these simple linguistic changes which generate a set of *perturbed* sentences from the dataset are called *perturbations*. In particular, in this work, eight different types of perturbations are generated ranging from punctuation deletion to various word-based transformations, i.e., substitution or insertion, preserving the semantics of the sentence. The purpose of these attacks is to make the model more robust with adversarial training. The way the approaches are evaluated to assess and improve the robustness of argument classification models is twofold: on the one side, the success rate of each perturbation type is evaluated on a model trained without any adversarial examples, and on the other side, the improvement in performance is measured on the original test set after augmenting the training data during adversarial training. As previously stated, the experimental setting relies on two standard datasets in Argument Mining, namely the *UKP Sentential Argument Mining Corpus* [119], and the *IBM Debater: Evidence Sentences* dataset [121].

Despite recent breakthroughs in modelling Natural Language Understanding, the employed neural architectures still lack interpretability. They are black boxes for which it is hard to determine what they exactly learn or are receptive for. In this context, it was found that Deep Neural Networks (DNN) are vulnerable to adversarial attacks; small changes to the input which fool the model into predicting a wrong label. Originally, crafting adversarial examples and attacking DNNs stems from the image processing domain [122–124]. Most of the employed methods there are gradient-based. These techniques cannot be easily adopted in the Natural Language Processing domain. Images consist of pixels, which are represented as real value vectors: it is possible to slightly change the pixel values in a way which manipulates the gradients in a forward pass of a model to change the prediction, while the image is still perceived as unchanged to a human. On the other hand, modifying a sentence in a way that a human will not notice that change is almost impossible. The main problem here is that while pixel values are represented in a continuous space, words – that can also be represented in a continuous space in the form of real value vectors, i.e., embeddings, – are in a discrete space per se. Theoretically, one could find a vector in the embedding space which changes the prediction of a model, but constructing this vector from a discrete space of words is impossible in most of the cases. So, the recommended option is to create a perturbation on a linguistic level in the target sentence. But, as said before, adding a word is most likely perceived by a human, contradicting the idea of an unnoticeable difference. Furthermore, adding even a single word might drastically change the semantics of

a sentence. Given these two challenges, adversarial examples in the NLP domain need to be carefully designed. Due to the nature of the problem, only limited work on the perceivability has been done so far. The main work focuses on semantic preserving techniques accepting that the perturbation might be noticed by the human eye [125].

A strategy to generate adversarial examples are black-box approaches. Contrary to white-box approaches, they do not need any model specific knowledge except the input and output. Recent black-box approaches comprise methods concatenating, editing or substituting words in the input sentence [125]. There are also approaches which work on changing the underlying syntax by creating paraphrases [126]. In the context of this thesis, I also experimented with this automatic paraphrasing technique to generate adversarial examples. While this is a highly interesting topic, for the argument classification datasets the produced paraphrases were ungrammatical most of the time. So, I decided not to further pursue this kind of perturbation and exclude them from the experiments. An intuitive way of creating perturbations is to replace words with semantically similar alternatives, e.g., synonyms. Alzantot et al. [127] employ an approach where they replace each word of a sentence until the prediction changes. For some of the presented perturbations, the same technique of replacing words with semantically similar alternatives is applied, but with a different strategy: only one word at a time is replaced minimizing the risk of producing a meaningless sentence. Moreover, also adverbs are added which change the semantics, strictly speaking, but do not change the label from argumentative to non-argumentative. Concerning the model which is attacked in the experiments, previous work has shown that self-attentive models are more robust than recurrent architectures [128]. While in this work the authors used a white-box approach to precisely aim at weak points of the self-attending model, I decided to pursue a model independent black-box strategy. The generated adversarial examples lay the foundation to evaluate the robustness of argument classification models and to improve it with adversarial training. Previous work on adversarial attacks in the AM domain [129] considers the Argument Reasoning Comprehension Task (ARCT) [113]. Here, given a claim, a reason and two warrants, the task is to determine the correct warrant. The authors found that the original dataset contains an uneven distributions of linguistic cues over the warrants, which the investigated model seems to learn [129], meaning that the correct warrant can be identified by learning these cues. To counter this imbalance, they created a perturbation for each data point (a claim C, a reason R, a correct warrant W and distractor warrant D). They negate the claim and invert the label (the identifier of the correct warrant) for each data point: For each $R \wedge (W \wedge \neg D) \implies C$, they add $R \wedge (\neg W \wedge D) \implies \neg C$ to the dataset. Since by the definition of the task R and W must be true to imply C, this is a logically correct transformation. While most of the negated claims already existed in the data, the remaining claims were negated manually. Contrary to the work proposed in this chapter, the perturbation is created manually and aims more at evaluating

the deeper comprehension of the logic of the argument and the learned inference than the flexibility of the model to handle linguistic variations of the input. Still, this is an interesting approach, showing that BERT cannot capture this deeper level of comprehension of arguments.

In the following, Section 7.1 discusses the methodology and background for adversarial attacks in NLP, with a subsequent focus on adversarial training on the argument classification task. The experimental settings are later detailed, including the used datasets and the generated perturbations, in Section 7.2, and the obtained results of the conducted experiments are discussed in Section 7.3. Subsequently, to round up the analysis, Section 7.4 highlights other observed weaknesses of transformer models.

7.1 Adversarial Attacks for Natural Language Processing

In this section, the terminology is introduced and an overview of the methodology for adversarial attacks on deep neural networks for NLP is given. For this work, I closely follow the definitions given in [124, 125] and explain which setting I chose for the topic-dependent argument classification task.

Perturbation: A perturbation is a minor change to the test input example for the DNN. The goal is to change the prediction of the model, while the modification of the input example should not be perceived by humans. As previously mentioned, the notion of being imperceptible by humans is not as easily applicable to text, because most of the time a change in characters or even words is more obvious to human judgment than a slight adjustment to pixel values. Thus, for NLP the point of perceivability is rather interpreted as preserving the semantics of the original sentence with being still grammatical as a further constraint. Both of these constraints are challenging NLP tasks by themselves and have not been fully solved so far. As a consequence, automatically generated perturbations might violate these constraints raising the necessity for a human evaluation of the generated perturbations.

Granularity of Perturbation: The notion of granularity follows the thought above. While slight changes in single characters might not be that perceivable and preserve semantics as well as syntax, deleting, inserting or replacing words is a different level of perturbation. Even changes on sentence level are possible, e.g., paraphrasing or even adding whole sentences as it was done for attacking reading comprehension models [130]. For the argument classification task, the majority of the perturbations are on word level, since the goal is to evaluate the robustness of the targeted DNN Language Model against comparatively simple linguistic attacks. Also one character-based method is employed and as aforementioned, unsuccessful experiments with sentence-level perturbations were conducted, i.e., automatic paraphrasing.

Adversarial Example: An adversarial example x' is a perturbation of an input example x, where the modification indeed changes the prediction Y of the model, so that $y' \neq y$.

Attack Target: An adversarial attack can be targeted to change only specific labels in a multi-class classification setting. For argument classification, there is no necessity to specifically target the attacks against a certain label for two reasons: first, argument classification is usually limited to a two or three class classification problem, and second I do not want to make any assumptions about the architecture of the attacked model, leading us to the next point.

Model Knowledge: There are different strategies to generate adversarial examples depending on the availability of knowledge about the DNN the attacks are aimed at. White-box approaches have access to all the information of the model, e.g., architecture, (hyper-) parameters, loss and activation functions, training data, or confidence scores. On the contrary, the black-box approaches have only access to the input and output of a model [131]. Everything between is unknown. Given that it was the best performing model in previous experiments (Chapter 4), BERT was selected as a specific model to attack. But since there are and will be other self-attending architectures based on Language Model pre-training, I do not want the perturbations to be limited to only BERT and decided to go for a black-box approach ignoring valuable information like the attention scores.

Adversarial Training: Currently, the only defense strategy against adversarial attacks is adversarial training where the DNN is re-trained with adversarial examples [122, 125]. One strategy is also to include inputs which are unlikely to occur naturally. This defense strategy aims at reducing the *"fundamental blind spots"* [123] of a model making the model more robust against divers input. With respect to NLP and specifically to argument classification, this means that including ungrammatical examples in training the model is justified. After all, argument classification is based on representations of full sentences, which are created from word level representations independent of the grammaticality of the sentence.

Evaluation Metric: The evaluation of adversarial attacks can be measured by the degree it decreases the performance of a DNN. I decided against it, because it cannot be ensured that each input example has the same number of generated perturbations, which thus might bias the results. Another prominent way to evaluate the perturbation efficiency is the success rate, which is used here as the evaluation metric. The success rate is the percentage of adversarial examples over the number of generated perturbations.

Robustness: In the terminology of this thesis, robustness refers to the ability of a model to correctly classify unseen test data from the same domain as the training data. Contrary to that, it is referred to generalizability as the concept of being able to exploit the already acquired knowledge in a new domain. For argument classification, this means that when training and test set talk about the same topics, e.g., *abortion*, adversarial attacks are testing robustness. For the case when the test set contains topics which are never seen during training, this falls under the (cross-topic) generalizability of a model. The main goal of adversarial training is to increase the robustness of a model, not its generalizability.

7.2 Experimental Setup

This section describes *i)* the datasets used for training and testing and the attacked DNN, *ii)* the different types of generated perturbations, and *iii)* a qualitative evaluation of the perturbations through a user study.

7.2.1 Data and Target Model

As previously mentioned, the application domain for the adversarial attacks in this work is topic-dependent argument classification. For this task, there are two major datasets available: 1) The *UKP Sentential Argument Mining Corpus* [119], which is a collection of 25,492 sentences annotated as an *ArgumentFor (**Arg+**), ArgumentAgainst (**Arg-**)* or *NoArgument (**NoArg**)* to a specific topic. The dataset comprises 8 different topics, i.e., *abortion, cloning, death penalty, gun control, marijuana legalization, minimum wage, nuclear energy* and *school uniforms*, and 2) the *IBM Debater: Evidence Sentences* [121], which is a collection of sentences from online debate portals annotated with *evidence **(Arg)*** or *no evidence **(NoArg)*** in regard to one of the 118 topics. Following existing experimental setups from the literature [38, 121], the training set comprises 83 topics (4,065 sentences) and the test set 35 (1,718 sentences).

Self-attentive transformer models like BERT [59], which use LM pre-training, have become a mighty tool for many NLP tasks. As stated above, this also applies to Argument Mining. Following recent state-of-the-art approaches to topic-dependent argument classification [38] and with respect to the experiments conducted for the previously described AM pipeline, the adversarial attacks were evaluated on the BERT_{base} model. The input for BERT consists of the input sentence concatenated with the topic. As introduced before, the perturbations are black-box methods not taking advantage of model specific knowledge, e.g., attention score, contrary to previous approaches on adversarial attacks on self-attentive models. Thus, they can be easily transferred to other architectures in the future.

Two lines of experiments were conducted. The first one to test the success rate of the perturbations, and the second one to evaluate adversarial training. For both lines, training and performance evaluation was based on the code provided by Reimers

et al. [38]. Hyper-parameters for fine-tuning the models were also replicated without any changes. The only difference is that I do not split the training data into a development set, since no parameters are tuned. For both lines of experiments, there are three different scenarios: 1) a model where the train (80%) and test (20%) sets comprise all eight topics of the UKP dataset (**UKP all**); 2) the leave-one-out training (**UKP x-topic**), where seven topics of the UKP dataset were used for training and the eighth is used for testing. In total, this results in eight different models. The results in this scenario are reported as the average over the eight models; 3) in the last scenario, a model is trained on the IBM dataset with the train-test split described above (**IBM x-topic**).

For the first line of experiments, i.e., perturbation evaluation, the success rate of a perturbation is evaluated on a model trained without any adversarial examples. Only perturbations from the test set are considered in calculating the success rate. For each perturbation, a label-wise success rate is calculated. For the second line of experiments, i.e., adversarial training, only perturbations of the training set are considered for augmenting the training data. Every model was re-trained under the same conditions as before, but with the only difference being the augmented training data. The evaluation of an adversarially trained model is done on the same unmodified test set as the normally trained counterpart to guarantee comparability.

7.2.2 Perturbation Types

In the following, the eight different methods are introduced, which were used to generate perturbations for given input examples. The perturbation generation methods are based on word or token types. Hence, the number of generated perturbations per input example varies. To give an idea of the order of magnitude, the average number of generated perturbations for each test set of the two datasets is reported.

Named Entities (NE) The first proposed method consists of replacing a named entity in the input sentence. To achieve this, a list of named entities is constructed for each of the four standard categories, i.e., *PER*, *LOC*, *ORG*, *MISC*, present in the CoNLL 2003 Shared Task dataset for named entity recognition [132]. Using this list, for each NE present in the original sentence one new perturbation is generated replacing the entity with a different entity from the same category. In order to preserve the semantics, pre-trained word embeddings (fastText) are employed as a means of distance, and the closest neighbours is selected. If the original input sentence does not contain a NE, no perturbations are generated. Accordingly, the average number of generated perturbations per input sentence varies. On the UKP dataset an average of 3.11 perturbations per sentence is produced. The IBM dataset contains more NEs per sentence, therefore the produced number of perturbations per example is higher, namely 10.15.

Example 7.2.1 *Original sentence: According to **FBI** statistics, 46,313 Americans were murdered with firearms during the time period of 2007 to 2011.*
*Adversarial attack: According to **U.S. Bureau of Investigation** statistics, 46,313 Americans were murdered with firearms during the time period of 2007 to 2011.*

Adjectives This method is similar to the list-based attack proposed by Alzanot et al. [127], where words in the input sentence are replaced with a word from a list of semantically similar words. Contrary to the aforementioned work, only one word per perturbation is replaced. Specifically, adjectives are exchanged with their synonyms, e.g., *big* with *large*, producing one perturbation example for each adjective in the sentence. The synonyms were taken from the WordNet interface in the NLTK. Here, the average perturbations generated per sentence are more similar in the two datasets. For the UKP dataset, a sentence has on average 2.12 adjectives, while for the IBM dataset 2.9 perturbations per sentence are generated.

Example 7.2.2 *Original sentence: A **big** part of it may have to do with the fact that marijuana today is much stronger than it was in previous generations.*
*Adversarial attack: A **large** part of it may have to do with the fact that marijuana today is much stronger than it was in previous generations.*

Punctuation This is the only modification of a sentence on character-level. Here, all the punctuation, e.g., "." or ",", is removed from the original input sentence. Naturally, this method provides one perturbation per sentence.

Scalar Adverbs This method is about adding or replacing emphasising modal adverbs, such as *considerably*, or trigger words for scalar implicature, such as *comparatively* or *largely*. They are added before a verb or an adjective. As will be shown in succeeding sections, the positioning algorithm needs to be improved, since some adverbs should be placed only after the word, while others should be placed only before the word or can take both positions. The average amount of perturbations generated per input sentences is around 3.94 for the UKP dataset and 4.67 for the IBM one.

Example 7.2.3 *Original sentence: It is possible to fuel nuclear power plants with other fuel types than uranium.*
*Adversarial attack: It is **totally** possible to fuel nuclear power plants with other fuel types than uranium.*

Nouns Similar to the proposed adjectives method, this list-based attack exchanges a noun with its hyponym. Again, only one word per perturbation is replaced producing one perturbation example for each noun in the sentence. This method generated an average of 12.19 perturbations per sentence on the UKP dataset, whilst the number increases to 17 for the IBM dataset.

Example 7.2.4 *Original sentence: When it comes to infertile couples, should not they be granted the* ***opportunity*** *to produce clones of themselves?*
Adversarial attack: When it comes to infertile couples, should not they be granted the ***chance*** *to produce clones of themselves?*

Conjunctions This method consists of adding adverbial conjunctions, such as *furthermore* or *nonetheless*, at the beginning of the input sentence. If the sentence already begins with an adverbial conjunction, the sentence is skipped. This attack delivers an average of 2.69 perturbations per sentence on the UKP dataset and 2.88 on the IBM.

Example 7.2.5 *Original sentence: Government data show that about one in 12 death row prisoners had a prior homicide conviction.*
Adversarial attack: ***Furthermore****, Government data show that about one in 12 death row prisoners had a prior homicide conviction.*

Speculative Adverbs They are modal adverbs related to the possibility property of verbs. This method is similar to the aforementioned scalar adverbs perturbation. Another list-based attack where modal adverbs related to the possibility property of verbs, such as *certainly*, are added directly before a verb. In this case, an average of 1.67 perturbations per sentence is obtained on the UKP dataset and 1.75 on the IBM.

Example 7.2.6 *Original sentence: Even the gateway effect — the theory that cannabis leads to other drugs — was discarded long ago.*
Adversarial attack: Even the gateway effect — the theory that cannabis ***indeed*** *leads to other drugs — was discarded long ago.*

Topic Alternatives Previous work has shown that including the topic in the BERT input increases the performance of the model [38]. Thus, exchanging the topic with alternatives is a relevant perturbation to evaluate. For each topic in the two datasets, a list of alternatives was created. For example, *arms limitation* for *gun control* or *capital punishment* for *death penalty*. While on average 4.25 alternatives per topic were created for UKP dataset, for the IBM dataset on average, there were 2.75 alternatives per topic.

7.2.3 User Study: Quality of Generated Perturbations

As an additional evaluation criteria of the generated perturbations, a user study was conducted about the preservation of semantics between the original sentence and the sentence after the modification. Both versions of a sentence were presented to the user and the user was asked if the two sentences 1) have the same meaning, 2) do not share the same meaning, or 3) if the transformed sentence is not meaningful, where "not meaningful" could mean either that the sentence has become ungrammatical or

that it does not make sense anymore. For each answer option, there was also a text field giving the possibility to voluntarily provide a justification of their decision. In total, 72 pairs of sentences were presented to each participant comprising every type of perturbation, but the topic alternative and punctuation deletion. The topic alternatives were excluded from the study, because the topic is an independent part of the model input and does not modify the grammaticality or semantics of a sentence. Same holds for the deletion of punctuation, which only changes the semantics of a sentence in some rare case of rhetorical questions. Moreover, the participant thinking of proper punctuation might have shifted their focus from the actual task, i.e., semantic similarity. The sentence length of each pair of sentences was controlled to have a difference of maximum one standard deviation from the mean sentence length of the sentences in the dataset. Participants in the user study were mainly non-native speakers with a higher educational degree (Master degree or Ph.D.) and a fluent level of English. In total, 31 people completed the questionnaire. The results are shown in Table 7.1.

Perturbation Type	**%**
Named Entities	71.30
Adjectives	61.04
Scalar Adverbs	42.67
Nouns	47.47
Spec. Adverbs	57.82
Conjunctions	93.68

TABLE 7.1: Results of the user study: percentage of how often each perturbation type was perceived as preserving the original meaning.

The perturbation method with the highest percentage of preserving the meaning of the sentence, i.e., 93.68%, is adding conjunctive adverbs. Naturally, this barely impacts the meaning of a single sentence. For the NE replacement, 71.3% of the people found the exchange as meaningful. The main criticism was that the new named entity, especially when they were acronyms, was unknown to the participant. Overall, employing word embeddings as a distance criteria to select NEs of the same type preserves the meaningfulness in most cases. Replacing an adjective with its synonym was in 61.04% of the cases found to be meaningful. While for the other cases, it was reported that the selected synonym was not suitable for the given context. Similar feedback was gathered for the hyponym replacement of nouns. Here, in 52.53% of the cases the selected noun did not fit the context, as either being too specific or unrelated to the topic. Inserting speculative adverbs was perceived as not changing the meaning of a sentence in 57.82% of the cases. A main observation reported by the participants is the change in credibility or certainty of the mentioned studies and other evidence, e.g., changing facts to opinions. Indeed, this does change the semantics of a sentence, but with respect to an argument classifier the uncertainty of an evidence does not matter as much as that it is correctly detected as being an argument. From this point of view, one can make an argument that this

aspect of change in semantics can be neglected in the particular case of argument classification, while it should be considered as changing the semantics in other AM tasks, where certainty plays a role. Compared with the other perturbation types, adding and replacing scalar adverbs caused with 57.33% the most cases of changes of a meaning of a sentence. The participants found that this transformation often breaks the grammaticality of a sentence. A future challenge is to find the right place to insert such adverbs, because some of them can either precede the target word or come only after it. Moreover, one has to consider if a target word can scale. For example, *genetic, mandatory* or *guilty* cannot be compared. There is no such thing as *fairly mandatory*. Future work in this research direction needs to address this point.

7.3 Results and Discussion

In this section, the results of the two lines of experiments are presented and discussed. First, the success rates for each perturbation type, and second, the adversarial training.

7.3.1 Adversarial Attacks

Table 7.2 reports on the success rate (the percentage) of adversarial examples over the total of generated perturbations.

Perturbation Type	**UKP all**			**UKP x-topic**			**IBM x-topic**	
	Arg+	Arg-	NoArg	Arg+	Arg-	NoArg	Arg	NoArg
Named Entities	7.06	7.30	2.02	6.14	7.22	2.30	1.51	0.18
Adjectives	10.90	10.02	6.70	12.16	10.37	5.89	3.79	0.03
Punctuation	8.86	9.74	4.21	10.41	10.61	4.34	2.78	0.19
Scalar Adverbs	5.87	7.15	3.41	7.39	7.57	3.29	2.01	0.08
Nouns	13.91	14.56	7.35	15.08	14.65	7.6	8.43	0.53
Spec. Adverbs	6.31	6.89	2.99	7.49	6.82	2.53	1.42	0.06
Conjunctions	5.87	7.29	4.33	9.66	9.52	4.56	3.64	0.4
Topic Alternatives	0.81	1.33	0.29	1.07	1.13	0.41	1.14	0.08

TABLE 7.2: Label-wise success rate of each perturbation type on the different test scenarios.

Looking at the in-domain test scenario, i.e., UKP all, one can observe that the Arg- label is more affected by the attacks than the Arg+ label, with exception of the adjectives. The adjective and noun replacement have the highest success rates in attacking the models. For adjectives, this could be explained with the fact that they usually carry sentiments whose perception might differ if they appear in a pro- or con-argument. For nouns, the replacement with hyponyms has the highest success rate, but given that in the human evaluation only in 47.47% of the cases the perturbation was perceived as meaningful, the results cannot be considered with respect to this perturbation as fully reliable.

Overall, the positive classes, Arg+, Arg- and Arg, showed to be more vulnerable to attacks than the no argument class. Usually, the structure of the task at hand,

which features in the data one tries to learn, is associated with the positive class. Meaning that the complementary class does not necessarily contain a distinctive pattern in the feature space, because it contains everything which is not wanted. Hence, it cannot be as efficiently attacked as the learned patterns for the positive classes. Unexpectedly, deleting the punctuation resulted in a comparatively high success rate, which is counter-intuitive, because one would not consider the punctuation to have a high impact on the model. This would translate to the model seeing meaning more in punctuation than actual meaning-carrying words, which is against any idea of Natural Language Understanding. And indeed, after reviewing the attention scores of the model, I found that, contrary to my expectations, the model tends to attend to punctuation. This further reinforces the suspicion that the model is not learning a deep semantic representation, because it does not comprehend the task as it was intended. The abstraction of finding correlations between the input symbols and the target labels to encode concepts and circumstances occurring in the world was not successful. This observation needs to be confirmed at a larger scale, though. Exchanging the topic with alternative wording resulted in an insignificant success rate not affecting the model. Concerning the cross topic evaluation, the UKP x-topic shows partially higher vulnerability than its in-domain counterpart. Since cross domain is the harder task, the confidence scores are lower for unseen test data, and with that the overall performance compared to in-domain models. A less confident model is easier to attack, explaining the higher success rates. Interestingly, the IBM x-topic is not as vulnerable to attacks as the UKP x-topic model. Again, as can be noticed in Table 7.3, the overall performance of the IBM model is higher. Since in both cases the same model architecture is employed, the only difference is the data. The IBM dataset seems to be more structurally uniform than the UKP dataset, explaining why test performance is higher and the success rate of attacks lower. Another point supporting this is that the exchange of NEs, which the IBM dataset contains more per sentence than the UKP one, barely changes the classification of an input example. This connotes that, in the case of the IBM data, NEs are not as important for the model justifying that they can be exchanged without losing the argumentative function of a sentence. Even though this further justifies the named entity perturbation method, it is ineffective in this case. Overall, BERT-based topic-dependent argument classification models are relatively robust against minor changes to the input, but still vulnerable to a certain degree. In roughly 5-10% of the cases, adding a meaning preserving word changes the prediction of the model.

7.3.2 Adversarial Training

The most common strategy to defend from adversarial attacks and make a model more robust is adversarial training. This is covered in the second line of experiments, whose results are reported in Table 7.3.

For the in-domain scenario (UKP all), one can observe an increase of .07 points in F_1-score compared to the model trained without adversarial examples. This shows

	UKP all	UKP x-topic	IBM x-topic
standard training	.73	.60	.77
adversarial training	.80	.59	.78

TABLE 7.3: Results in macro F_1 for models with and without adversarial training.

that adding linguistic variants of the training data helps in predicting unseen test data from the same domain. Intuitively this makes sense, arguments are often rephrased differently or are re-used as targets for undercutting, for example. With respect to BERT, this raises questions. In the aforementioned experiments on perturbation efficiency, it was seen that BERT seems to be quite robust against the adversarial attacks. Also, in previous works, models based on Language Model pre-training advanced the state-of-the-art, which was said to be due to the Natural Language Understanding capabilities learned during pre-training. Accordingly, this should mean that slight variations of the input are covered by the Language Model. The increase in performance with adversarial training shows that this supposed NLU capability is either not fully utilized or blurred during fine-tuning, or was limited in the first place. I assume it is a mixture of both, since other experiments in different domains show that BERT-like models are more robust than recurrent networks [128], but also that the Language Modelling capabilities of self-attentive models are limited [133, 134]. Even if the success rates of the perturbations are only between 5-10%, added up these make quite a number of examples, which BERT is vulnerable to. Adding these linguistic variations to the training data, though, boosts the NLU capabilities making the model more receptive for them. Note that this way the training data is increased by roughly a factor of twenty. This indeed shows that adversarial training helps in-domain predictions and improves the robustness of a model, as intended. Table 7.4 shows examples where adversarial training corrected the model prediction.

topic	**sentence**	$pred_1$	$pred_2$
gun control	Five women are murdered with guns every day in the United States.	NoArg	Arg+
school uniforms	Up to now, this uniform is still in use, making it the oldest uniform in history.	Arg+	NoArg
abortion	Even in the case of nonfatal conditions, such as Down syndrome, parents may be unable to care for a severely disabled child.	Arg-	Arg+
cloning	I find this reasoning absolutely ridiculous, since a person is a person despite their genetic source or if artificially created.	Arg-	Arg+

TABLE 7.4: Examples where adversarial training improved the model prediction. $pred_1$ model prediction before adversarial training, $pred_2$ model prediction after adversarial training, which is also the true label.

A justified doubt coming up here is the question of overfitting, i.e., *did the adversarial training really help in NLU or did it just improve learning the dataset?* In the latter case, one would see a decline in cross domain evaluation, because the model is overly focused on in-domain specific features. As can be seen in Table 7.3, the cross domain performance is not dropping significantly with adversarial training. Both models are still in an acceptably similar range compared with their normally trained counterpart. The UKP x-topic loses 0.01 F_1-score, while the IBM model even shows a slight increases of roughly .01 F_1-score. Meaning that the generalizability of the models is preserved, ergo they did not overfit on the training domain. So, *why is it that adversarial training helps in-domain, but does not improve the cross domain performance?* At this point, we need to go back again to the aforementioned distinction between robustness and generalizability. On the one hand, robustness is more related to the ability to understand language in the sense of linguistic flexibility; being able to understand differently worded phrases about the same thing. Generalizability, on the other hand, is the ability of a model to transfer and apply already learned patterns to a new domain. In this case, an increase in performance for the models tested on cross topics would be related to the generalizability. While depending on the application scenario, generalizability and robustness have a strong overlap, and one has to carefully distinguish them for Argument Mining. Usually, cross domain in AM means that the model should be able to detect arguments for a topic unseen during training. Assuming the new topic is not somehow related to the topics seen during training, this means, the model has to infer everything associated with a given input sentence and decide if this can be an argument related to the topic or not. The problem is one can only conditionally infer new arguments from existing arguments in the semantic space. If the two arguments are structurally similar to a certain degree (or use similar key components), it is possible. But finding new arguments for an unseen domain is beyond Language Modelling. It requires also a deep understanding of knowledge and common sense. Especially the latter two cannot be efficiently learned from word co-occurrences alone [118, 134]. As a result, it is not surprising that augmenting training data with alternative wording of the data does not improve generalizability. After all, the examples added for adversarial training are mostly noise with respect to the new unseen test domain; noise, which is not negatively affecting the generalizability of the BERT model.

7.4 Known Weaknesses of Transformer Models

As described in Chapter 2.3, a whole field of research has been developed analysing the inner functioning of the attention mechanisms of transformer models [135]. This goes far beyond than just the evaluation of robustness. This research has discovered various weak spots of BERT-based models. In the following, various points together with proposed improvements are presented.

Shortly after the work I presented in this chapter, Ribeiro et al. [136] proposed a task-agnostic methodology for testing NLP models. There, NLP models should be tested for certain changes in the input, i.e., perturbations. In line with my findings, they found that current transformer models struggle with changes of locations and person names. They also evaluated the vulnerability to paraphrasing and discovered that contemporary commercial transformer models are far way from solving this problem. Moreover, they cannot capture temporal changes of verbs, e.g., changing *is* to *used to be*, and negations, where the latter was also observed in the error analysis of the relation classification and Effect-on-Outcome in the Chapters 4.2 and 6.3.

Contemporary transformer models surpass the performance of non-expert humans making the GLUE benchmark no longer a suitable metric to reflect the improvements models make [137]. The newer SuperGLUE [137] benchmark comprises harder tasks like reading comprehension, common sense reasoning or textual entailment to better quantify the performance of the understanding. While for most tasks the leaders of the GLUE benchmark also performed reasonably well [133], they are significantly worse than humans on the causal reasoning task [138] and co-reference dependent reading comprehension [139], where the human baseline is at 100% accuracy. Besides a deep understanding of the discourse, these problems require common sense and world knowledge. It has been shown that BERT-based models in the higher layers do capture some kind of semantic abstraction [73] and the performance of the models on the aforementioned tasks is also high. But some questions arise, e.g., how well do these models understand the interactions in a discourse and how much common sense and world knowledge can be learned from just word co-occurrences? and more importantly, how can the major limitation of being trained only with character-based features be enhanced to capture more of this information? One option is to include semantic information in the training process, where semantic information can either mean world knowledge from knowledge bases or integrating discourse and semantic role information. Additionally to the aforementioned problems of understanding the discourse, knowledge dependent tasks, like fine-grained relation classification or entity typing, pose challenges for models trained solely on contextual character-based features. For example in *Bob Dylan wrote Blowin' in the Wind in 1962*, it is hard to determine if Bob Dylan is a writer or songwriter without knowing that *Blowin' in the Wind* is a song. This knowledge is available in Knowledge Bases (KB). The semantic web is full of structured world knowledge, which can be exploited. One approach to incorporate such external knowledge into Language Models is *Enhanced language RepresentatioN with Informative Entities (ERNIE)* [72]. The idea is to stack a knowledge encoder consisting of multiple aggregators on top of the encoder layers of a transformer model, where the knowledge encoder fuses knowledge graph embeddings with the contextualized embeddings into one united feature space. As a first step, named entity mentions in the text are aligned with their KB entries. The aligned named entities are represented

with knowledge graph embeddings using TransE [111]. Each aggregator takes the contextualized token embeddings from the transformer encoder and the entity embeddings and feeds them into a multi-head self-attention layer, respectively. An information fusion layer integrates the different representations coming from the two self-attention layers into one feature space. The output embeddings for each token and entity are the input for the next aggregator. The output of the last aggregator is used as the final embedding representation. For more details I refer the reader to the original paper [72]. Like BERT, the pre-training for ERNIE is done with cloze test like tasks[1]. Similar to the masked Language Model, they employ a knowledge masking task, where either one entity of the entity alignment is replaced with a random entity, a token-entity alignment is masked, or the alignment stays unchanged. For ERNIE 1.0 the pre-training comprises MLM, next sentence prediction (same as for BERT) plus the knowledge masking task, while ERNIE 2.0 consists of more tasks. Adding only the knowledge masking to the pre-training, ERNIE 1.0 significantly outperforms BERT on entity typing and relation classification datasets while still delivering comparable results on GLUE. With ERNIE being a first step towards integrating heterogeneous information coming from world knowledge databases, the next step is to inject common sense knowledge in a similar fashion. There are available resources providing this knowledge to a certain extend, e.g., ConceptNet [140] or ATOMIC [141] in form of cause and effect relations. Moreover, in direct relation to the work of this thesis, integrating world knowledge or common sense could mean explicitly modelling the implicit warrants, as discussed in Chapter 4.2.

One approach to include contextual semantics to Language Modelling is SemBERT [70], motivated by the semantically incomplete answer spans of BERT on the Stanford Question Answering Dataset (SQuAD), where single semantic discourse units were broken down and only parts were classified as the answer to the question. A similar problem was observed for incomplete outcome span detection in the outcome analysis pipeline in Chapter 6.3. A problem shown for SQuAD was, for example, answering *How many people does the Greater Los Angeles Area have?* with *17.5 million* instead of *over 17.5 million*. To overcome this problem, the authors integrated information from Semantic Role Labeling (SRL) in the sequence encoding. As a first pre-processing step, the input sentences are annotated with a semantic role labeler. Each token is assigned a list of labels, where the length of the list is the number of semantic structures output by the semantic role labler. The embeddings of each semantic role label are learned via a BiGRU and subsequently fed into a linear layer to obtain one joint representation for each word in the sequence. In parallel, the subword level representations from the BERT encoder are converted to word-level using a CNN with max pooling to match the token length of the SRL output. The contextualized and semantic embeddings are concatenated to form the final embedding. While the BERT encoder is initialized with pre-trained weights, the weights for the

[1] As explained in Chapter 2.3, cloze tests are fill-in-the-blank tests, which require an understanding of the context and are commonly used in language learning.

BiGRU are learned during fine-tuning on a specific task. SemBERT outperformed the existing models on GLUE and SQuAD[2].

Another way to inject discourse knowledge is discourse-aware semantic self-attention [71], which replaces the basic multi-head self-attention block in the transformer encoder. Here, the motivation comes from integrating discourse information into reading comprehension to better understand interactions, causation and temporal sequences in the text. For example, given the context: *Jacob frequently visits Jeff and Kenny, who are serving time in a juvenile hall. Jacob initially threatens them, until eventually Jeff commits suicide.* To answer *Why did Jeff commit suicide?* one needs to understand that the suicide is caused (*until eventually*) by Jacob threatening Jeff (*them*). For this, structured knowledge about entity co-reference and their semantic roles are required as much as information about the discourse relations between text sequences. To learn all this information, the proposed self-attention gets three additional inputs[3], which are represented by one embedding vector, respectively: 1) semantic role label; similar to the aforementioned approach, embeddings for the semantic roles are learned. 2) discourse relation label; following 15 fine-grained discourse relation sense types from the Penn Discourse Tree Bank annotation scheme, such as *causation* or *contrast*. 3) label of the co-reference cluster; where tokens referring to the same entity are assigned to the same cluster. Using these linguistic annotations, the model outperforms the same model with the basic self-attention by +3.43 Rouge-L on NarrativeQA reading comprehension. Concerning the impact of the individual linguistic information, the authors found that information about the SRL improves *who* and *when* questions, while information about the discourse relations is beneficial to answer *why* and *where* questions. *Why* questions in particular are relatively close to AM tasks, which further supports the integration of discourse information in AM models as it was done in [142], who showed that discourse parser features can contribute in argument parsing.

Similar to the discourse-aware semantic self-attention, ERNIE 2.0 [69] takes advantage of information about the discourse relations. One of the added tasks for pre-training with respect to the previous version, is the discourse relation classification task. Here, the model has to predict the marker, e.g., *but*, for an explicit discourse relation between two sentences. Together with the continual learning strategy and the other added pre-training tasks related to lexical, structural and semantic information, ERNIE 2.0 shows significant improvement compared with the previous version.

Therefore, I consider the addition of semantics to LMs trained on only contextualized character-based features an important and inevitable step towards Natural Language Understanding and AM in particular. Especially with respect to common

[2]While later models like XLNet and RoBERTa outperform SemBERT, they still do not consider semantic information. The proposed approach to inject semantics can be implemented in these LMs as well.

[3]Linguistic annotation is a pre-processing and relational annotations spanning multiple sentences are projected from paragraph-level to token-level.

sense, world knowledge and co-referential discourse, current contextualized representations cannot solve the challenges of general language understanding alone. The latter, as was discussed throughout the preceding chapters, is decisively important for AM, since it is such a highly complex problem. Again, a shallow understanding based on symbol patterns is not sufficient to fully solve the challenges posed by the definition of the problem. The task is to encode concepts and circumstances occurring in the world in such a way that the (Machine Learning) model can infer the causal relationship between two argumentative statements and is able to transfer this learned knowledge to leverage it on new data. This requires a deep semantic understanding of both statements, their interactions and following consequences. As I said previously in this chapter, I think that the abstraction of the current (transformer) models does not reach fully into a deep semantic space. They are limited to a mapping of input symbols into a predefined space and learning the distance units and *meaning* of this space on their own. With respect to the context of this thesis, especially in the medical domain, interrelationships are often implicitly presumed and not explicitly mentioned in the text. This world knowledge has to be externally added to the model, because it is impossible to capture with a model trained solely on character-based input in an unsupervised way. In my opinion, an approach without any additional information about the discourse, common sense or world knowledge, cannot capture all properties of communication. Language is after all just a tool to transfer information based on observation, consensus and experience in social interaction of the speakers. All these factors play an important role in arguing and debating, which makes them relevant for the tasks of AM and should therefore not be ignored. Thus, I agree with Niven and Kao [129], that the current transformer models are strong learners of linguistic cues, but cannot solve AM tasks satisfactorily beyond a certain point, since they simply cannot comprehend all facets of the argument(-ation), yet.

Chapter 8

Proof-of-Concept and Impact

This chapter introduces the ACTA system and its applications in concrete healthcare scenarios linked to the Covid-19 health emergency. ACTA is a tool for automatically analyzing clinical trial abstracts from the argumentative point of view by finding argument components and their links. Moreover, PICO elements are detected and highlighted. In the context of the Covid-19 pandemic, ACTA was updated including the extension of the relation classification and it has been employed in the Covid-on-the-Web project. Furthermore, the output is stored as RDF data through the use of ontologies for data representation. Within the project, ACTA is integrated in the overall data processing pipeline to create Linked Data. This chapter comprises the work published at the International Joint Conference on Artificial Intelligence (IJCAI-2019) [143] and the International Semantic Web Conference (ISWC-2020) [144].

In Chapter 2.1 it was said that one aspect is that Argument Mining can assist in scrutinizing the conclusions drawn by the authors of a trial. Creating an argumentative representation of the trial(s) can support clinicians and practitioners in interpreting the results and take informed decisions. The other scenario where AM could help is based on the rising popularity of argument-based decision making in medicine, as discussed in Chapter 9.1.1, where AM extracts the structured data from unstructured text required for these types of decision support. Both are eminently application oriented objectives. Thus, to demonstrate the feasibility and benefit of the proposed AM pipeline on clinical trials, a demo system was developed, called **ACTA**[1]. ACTA stands for ***A**rgumentative **C**linical **T**rial **A**nalysis* and is, as the name suggests, a tool to automatically analyse the argumentative information of clinical trials. It may be seen as the first step of a pipeline ending with evidence-based decision making frameworks in healthcare applications, as those illustrated in Chapter 9.1.1. The main purpose of it is to support the decision making process in EBM, by visualizing trial abstracts. The displayed summarized information about PICO elements and contained arguments should facilitate literature exploration. Hence, ACTA integrates the web interface for literature research with PubMed, which allows the

[1] http://ns.inria.fr/acta/

user to query biomedical literature as usual. After the query the user can choose the documents to be analysed and run the pipeline on the selected articles. The argumentative components are then automatically extracted and the links between the components established. In the base version of ACTA, the argument structure prediction does not comprise the relation classification and only unlabeled links are predicted, where the task is formulated as a multiple choice problem, as described in Section 4.2. Besides the argumentative information, ACTA extracts PICO elements from the trial abstract. As mentioned several times, PICO elements play an important role in EBM, especially as a source of information to appraise literature. Hence, adding a PICO element detection module to ACTA was an essential step towards adapting the AM pipeline for the needs in the medical domain. The PICO element detection module provides information about the mentioned participant, intervention and outcomes of a study. As a hybrid tool, ACTA conveys information in a form the medical user is familiar with and might thus be more likely to be accepted. After all, the objective of ACTA is to give a condensed yet valuable overview of clinical studies to assists in the deliberation process.

At the moment of writing, a potential application of ACTA by Inserm[2] is discussed. Inserm (***I**nstitut **n**ational de la **s**anté **e**t de la **r**echerche **m**édicale*) is the French National Institute of Health and Medical Research. It is the number one applicant of patents in Europe in the pharmaceutical sector and takes the second place in the SCImago Institutions Rankings[3] for best research institution in the health sector behind the National Institutes of Health in the United States[4]. As a prestigious public scientific and technological institute it is involved in an entire range of activities resulting alone in 2019 in 11,700 publications[5], which are 36,5% of all biological and medicine papers worldwide. To assess the publications of their researchers and have a better overview of what their research is about, ACTA is discussed as a potential tool. While this application scenario goes beyond decision support for EBM which I had in mind when developing ACTA, it is an unforeseen but worthwhile use case. This further demonstrates the versatility and practicality of AM techniques in the broader context of (medical) research.

In Section 8.1, the original base version, presented as a demo at the International Joint Conference on Artificial Intelligence 2019 [143], is introduced. In Section 8.2, a first updated version with additional features, which was applied on the CORD-19 dataset [145] in the context of the Covid-on-the-Web project [144], is presented together with the objectives of Covid-on-the-Web project itself.

[2] https://www.inserm.fr/
[3] https://www.scimagoir.com/
[4] https://www.scimagoir.com/rankings.php?sector=Health&year=2019
[5] https://www.inserm.fr/en/about-inserm/inserm-glance

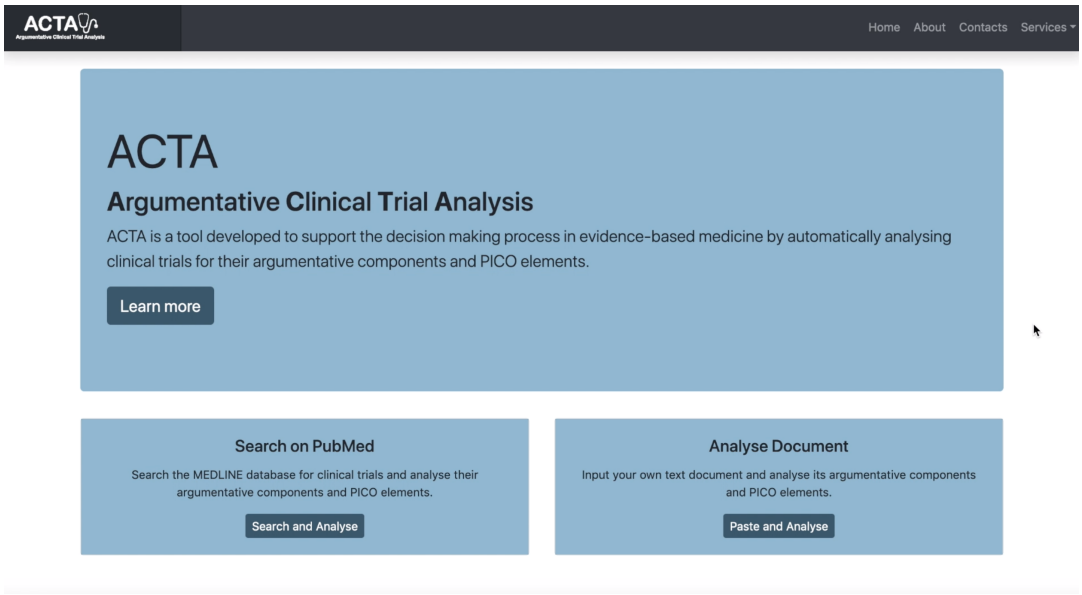

FIGURE 8.1: The ACTA main page.

8.1 ACTA

The ACTA tool [143] is designed to support doctors and clinicians in identifying the document(s) of interest about a certain disease, and in analyzing the main argumentative content and PICO elements. ACTA automatically analyses the textual abstract(s) of clinical trials that the user provides, and it detects in the text the argumentative components, i.e., evidence and claims, together with their relations. In addition, the identification of PICO elements in the abstracts is included. ACTA returns the user with the argumentative structure identified in the selected abstract(s), under the form of a navigable graph whose nodes are the argumentative components. PICO and argumentation elements are highlighted in the textual abstract with different colors.

The main features are illustrated in detail in Section 8.1.1. ACTA employs Argument Mining methods to identify the argumentative structure of textual clinical trial abstracts, which are described in Section 8.1.2.

8.1.1 Main Features

ACTA goes beyond the basic keyword-based search in clinical trial abstracts, and it empowers the clinician with the ability to retrieve the main claim(s) stated in the trial, as well as the premises (or evidence) linked to this claim. As a result, the clinician does not need to read the whole abstract, but is provided with a structured "summary" of the abstract under the form of a graph. More precisely, ACTA provides clinicians with the following facilities: Search options for PubMed, a custom text input option, the argumentative analysis together with the PICO element detection, and an option to download the results in form of a `json` file. Figure 8.1 shows the ACTA main page, which is the entry point for the user.

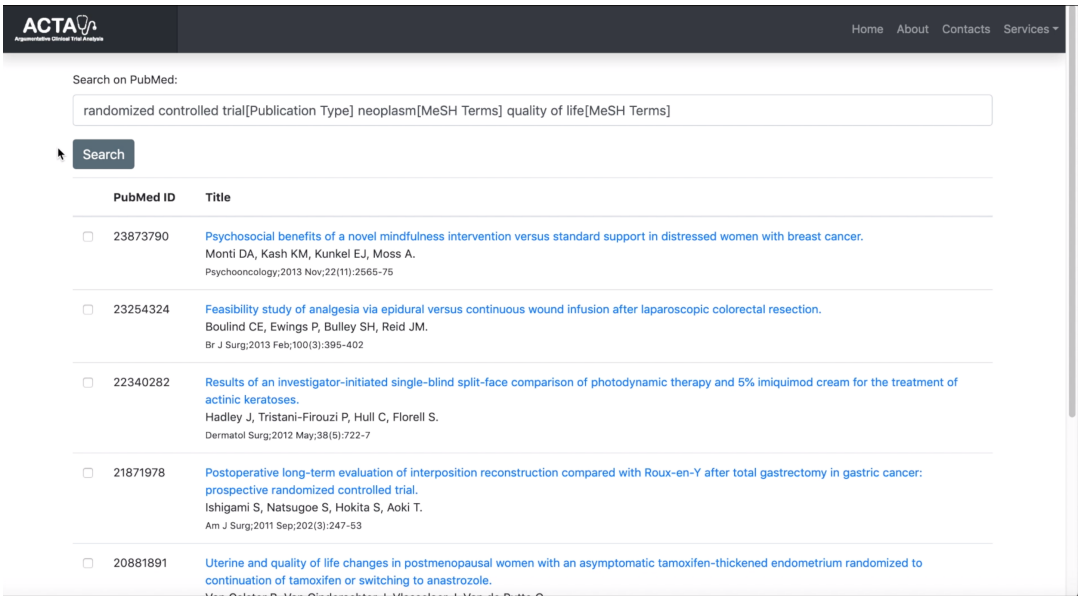

FIGURE 8.2: Illustration of PubMed search interface in ACTA.

Search on PubMed

PubMed[6] is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. Given the importance of this search engine in the healthcare domain, the possibility to search for a (set of) abstract(s) directly on the PubMed catalogue is included in ACTA, through their API. This way, the mode of inquiry is unchanged and the user can keep using the familiar terminology, including MeSH[7], to express the query similar to the advanced search builder from PubMed. The interface is shown exemplary for the query *"randomized controlled trial[Publication Type] neoplasm[MeSH Terms] quality of life[MeSH Terms]"* in Figure 8.2. This query pattern was also used to extract the RCTs for the second version of the AbstRCT dataset, as mentioned in Chapter 3. It specifies that the clinical study has to be a Randomized Controlled Trial about interventions for neoplasm with quality of life being one of the measured outcomes. The *AND* connectors between the single search parameters are added automatically by calling the API. *OR* connectors have to be specified manually.

After the search is executed, the results are listed below the search bar. The contained information for each entry comprises all the relevant information, such as the PubMed ID, authors and publication date, as provided in a search directly on PubMed. Moreover, each result directly links to the document entry page in PubMed. When the search results are shown, the user can select one or more abstracts to address the argumentative analysis. As an alternative to the PubMed search, the user has the option to directly enter a trial abstract or other text in an input field to get analyze with ACTA.

[6] https://pubmed.ncbi.nlm.nih.gov/

[7] As mentioned earlier, MeSH is the vocabulary thesaurus used for indexing biomedical articles.

Argumentative Analysis and PICO Elements

As soon as the free text is entered or the abstract(s) is selected from the PubMed search result list, the user can run the argumentative analysis by pressing the analyse-button. This will forward the user to the result navigator and the visualization of the abstract. In a panel on the left side, all documents, which were selected in the preceding step in the PubMed search, are listed with their PubMed IDs (PMID). This can be seen in Figure 8.3, where five documents were selected to be analysed from the research results. By clicking on one, the respective document is visualized. The result is visualized to the user under the form of an argumentative graph (middle of the window) where the nodes are the evidence and the claims automatically detected in the abstract, together with their links. The nodes are annotated with their type and ID. The textual content of the argumentative component is shown, when the user hovers over a node, this is illustrated in Figure 8.3. In addition, the full text of the abstract is shown on the right side of the graph together with other meta information about the selected document, i.e., the PMID, title and authors. There, the user also finds the download-button. By clicking on it, the currently visualized document is downloaded with all annotations in a `json` file. Furthermore, under the displayed abstract, the user finds the options to highlight evidence and claims with different colors in the abstract. This is shown in the upper screenshot in Figure 8.4, where evidence are marked in yellow and claims in blue. The highlighting colors match the colors of the nodes in the argument graph. Besides the highlighting of the argumentative components, the detected PICO elements can be accentuated, as presented in the lower screenshot in Figure 8.4. There, the found participants/population of a study are marked in green, e.g., *patients with advanced epithelial ovarian cancer*. All interventions (including the comparison intervention) are colored in red. As explained in Chapter 6, for practical reasons, it is not distinguished between intervention and comparison intervention. Sticking with the example in Figure 8.4, this means that both, *platinum plus paclitaxel* and *gemcitabine*, are highlighted in red. The final PICO element, the outcomes (*progression-free survival, overall survival* and *objective response* in the case of the example in Figure 8.4), are accentuated with a purple background. The user can switch between both illustration modes with the highlight-buttons. Additionally to this pictured information, both types of results, the argument components and the PICO elements, are listed as tables below, depicted in Figure 8.5. This gives a more structured way of showcasing the information, which comes in handy specially for the PICO elements. At the current stage, the PICO elements are listed in the order they occur in the abstract. They are not yet filtered and duplicates or very similar expressions may be listed. For example in Figure 8.5, *Topical photodynamic therapy (PDT) with aminolevulinic acid (ALA) and 5% imiquimod cream* and *ALA-PDT and imiquimod 5% cream* are the same intervention, but are listed separately. Future work could try to develop a method to merge similar elements into one expression. First rudimentary steps towards this were taken, as described in the Section 8.2, but are still far away from grouping together related

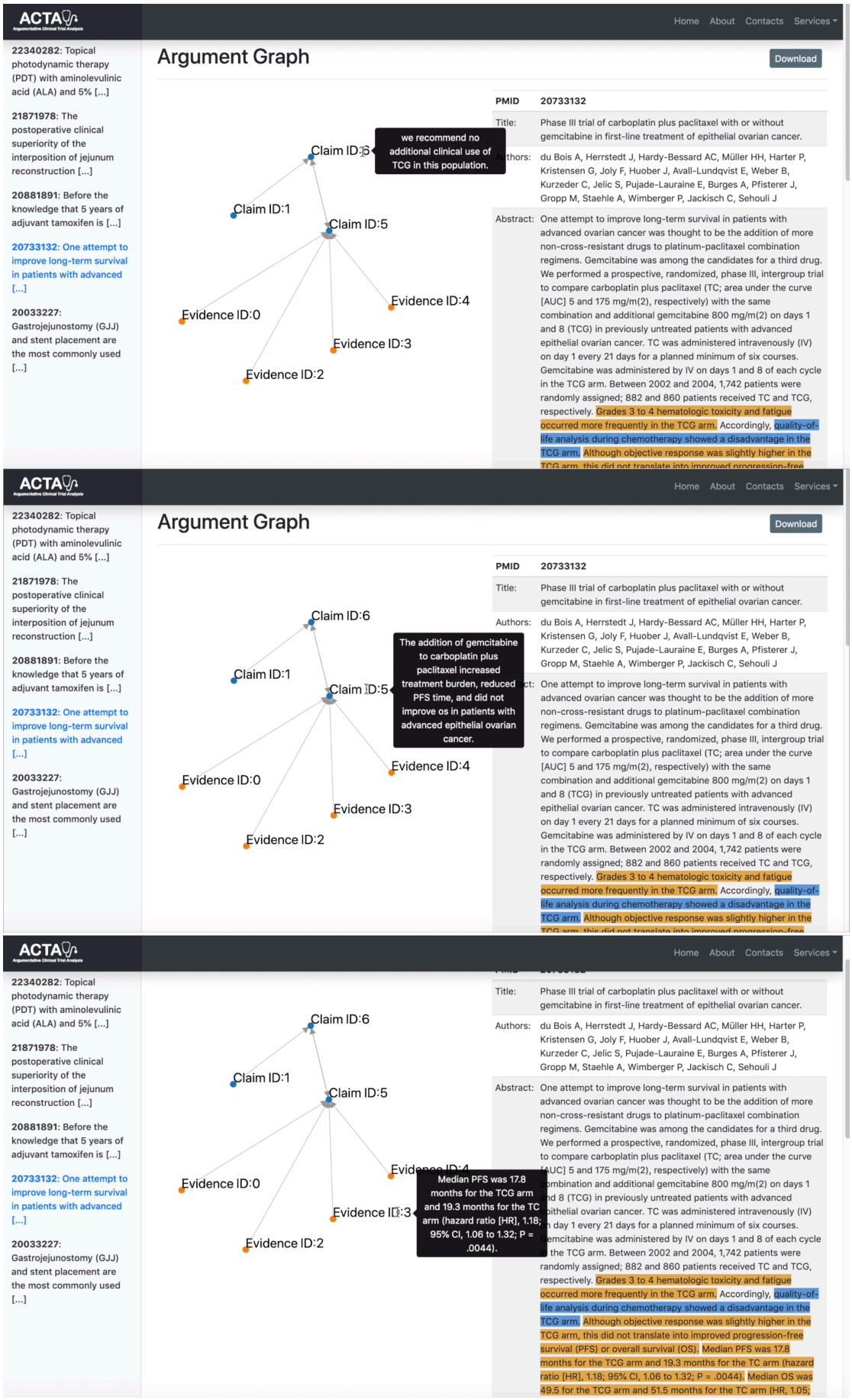

FIGURE 8.3: Multiple screenshots to illustrate the different functionalities of ACTA and the visualization of the argument graph returned to the user.

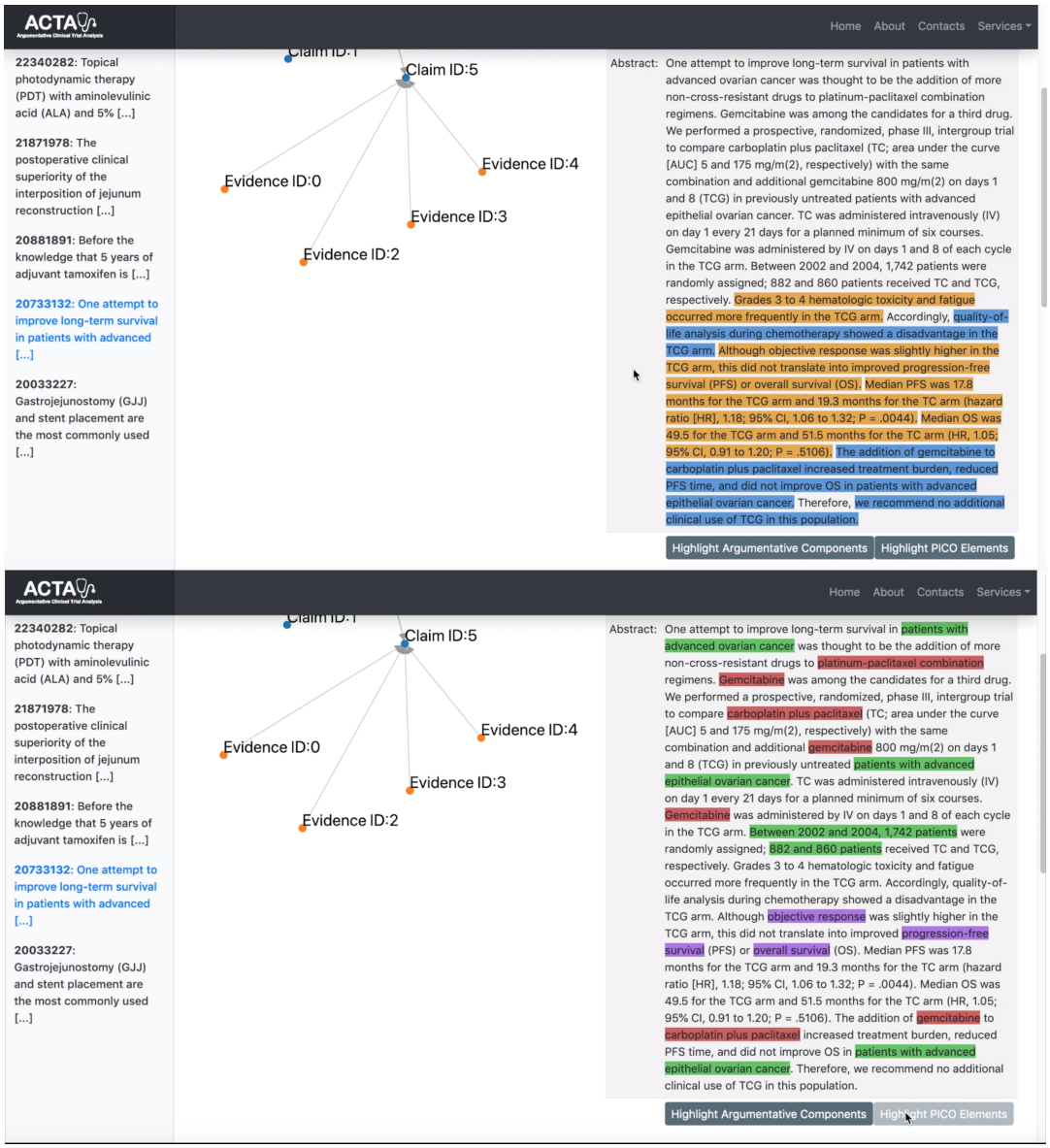

FIGURE 8.4: Screenshots showing the highlight options for the analysed document in ACTA.

or identical elements successfully. To give a reference of the difficulty of this task, the authors introducing the used dataset of PICO elements report results lower than majority voting for the detection of redundant information [6]. Especially combined treatments, such as ALA-PDT, are challenging to discriminate from other combined treatments, e.g., ALA-PDT plus imiquimod.

8.1.2 Experimental Setting and Results

For the argumentative component classification and boundary detection, the AbstRCT dataset with 500 abstracts of randomized controlled trials on neoplasm treatments was used. The relation annotations are used for the link prediction task in the

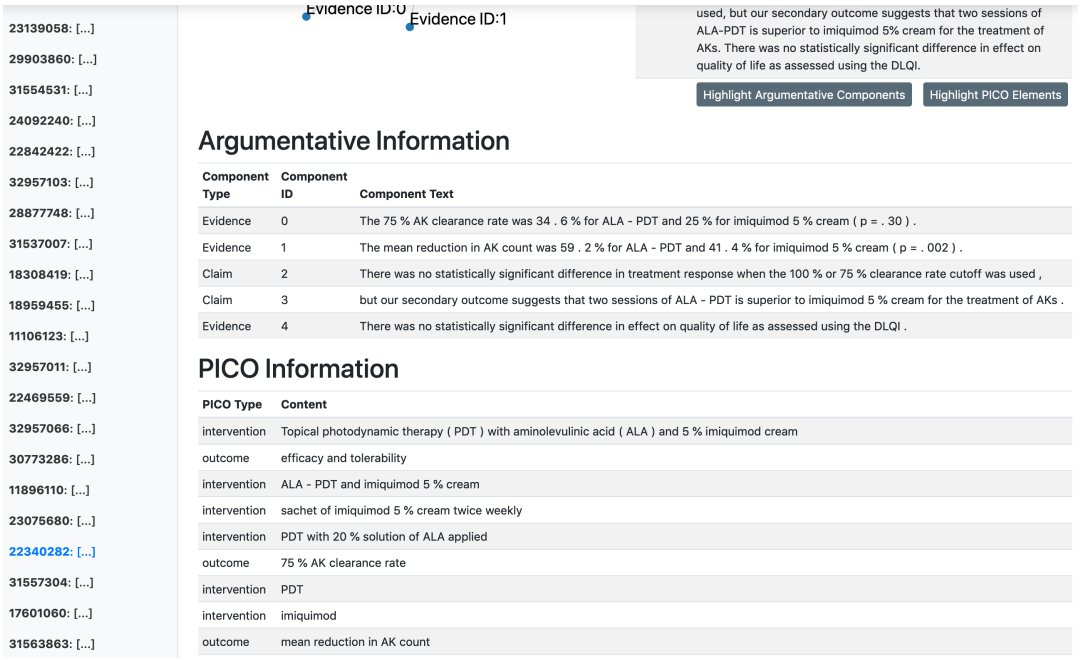

FIGURE 8.5: Screenshot showing the displayed argumentative and PICO information in ACTA.

base version of ACTA. In the later updated version, described in Section 8.2, the relations types, i.e., *Attack* and *Support*, are considered, too. Since ACTA was developed shortly after the first experiments with the BERT$_{base}$ model, it resembles the experimental design described in Chapter 4.1.2: The argument component detection is treated as a sequence tagging problem with the BIO-tagging scheme. At the time of development the SciBERT pre-trained weights were not yet published. Thus, the applied model is the BERT$_{base}$ model with a shallow layer for sequence tagging, which in this case consists of a BiGRU and a CRF. The entire model is fine-tuned achieving a macro F_1-score of .85 on the neoplasm test set, as reported in Chapter 4 Table 4.3.

The same method is applied to train the model for the PICO element extraction. As data, the EBM-NLP dataset [6] with coarse labels is used. Coarse labels are *population, intervention* and *outcome*. The fine-grained labels naturally provide more information, for example, *Age* and *Sample size* are more expressive than just *population*. On the other hand, the outcome distinction between physical health and mental impact, seemed too particular for the application scenario with ACTA. Moreover, they are harder to learn. The creators of the dataset report a difference in performance of .40 F_1-score with their baseline model (BiLSTM+CRF) in the fine-grained labeling task for participants and interventions. This decrease in performance would drastically reduced the practical use of the demo system. Thus, since PICO element detection is not a trivial task, to find the right balance between performance and information content, the coarse labels were selected. Accordingly, the model was trained to jointly predict the participant, intervention and outcome candidates for a given input. Dataset splits were the same as in the original paper, with the difference that sentences containing less than 10 WordPiece tokens [58] were ignored. The BERT$_{base}$

model achieves .73 F_1-score on the test set. Some of the challenging cases and errors can be seen in the lower screenshot in Figure 8.4. For example, the misclassification of date spans as reports of population: *between 2002 and 2004*. Further, specifications of doses of drug interventions, e.g., *gemcitabine* ***800 mg/m(2)***, which are an important detail, cause complications and are not always added by the classifier.

Regarding the prediction of the links between argumentative components, it is treated as a multiple choice problem, similar to the *Situations With Adversarial Generations* (SWAG) task [108], where one has to select the correct target sentence for a sentence-pair from a list of possible candidates. This way it is ensured that each source component has a maximum of one link to a target component. I considered this important for the targeted argumentation graph, which allows one outgoing edge per node at most to eschew divergent argument structures. As seen for the SentClf approach in Chapter 4.2, divergent structures are most of the time false positives, but not as common as expected. For training, the multiple choice BERT_{base} model, compare with Chapter 4.2, is fine-tuned for three epochs with an Adam optimizer and a learning rate of 3e-5, resulting in .79 F_1-score for the binary evaluation of the link prediction. As explained in Chapter 4.2, the multiple choice model was later developed further to include the relation classification.

At the current stage, ACTA is only apt to analyse English documents. Given the lack of AM datasets in the medical domain in different languages, no training data is available for other languages. This could be bypassed with zero-shot learning attempts. However, this still requires a multilingual test set for evaluation, which is at the current moment not available. Nevertheless, the feedback from exhibiting the demo system at the International Joint Conference on Artificial Intelligence was promising and encouraged me to deeper entwine AM methods with elements relevant for EBM, as it was done, for example, with the Effect-on-Outcome in Chapter 6. Other extensions and improvements of the pipeline, which were left for future work in the published paper of the base version, were addressed in the context of the *Covid-on-the-Web* project, which is described in the next section.

8.2 Covid-on-the-Web Project

With the Coronavirus infection disease (Covid-19) spreading in the spring of 2020, the Wimmics research team[8], where I am part of as a doctoral student, decided to join the effort of many scientists around the world who harness their expertise and resources to fight the pandemic and mitigate its disastrous effects. A new project, called *Covid-on-the-Web*, was initiated with the goal to facilitate the access, querying and information processing of COVID-19 related literature for biomedical researchers. To this end, tools were adapted/re-purposed and combined to publish, as thoroughly and quickly as possible, a maximum of rich and actionable Linked

[8] https://team.inria.fr/wimmics/

Data about the coronavirus. Since it was a user-oriented project, the main motivating scenarios were identified through a need analysis of the collaborating biomedical institutions, i.e., the French Institute of Medical Research (Inserm), the French National Cancer Institute (INCa) and the Antibes and Nice Hospitals. The main scenarios addressed with the provided Linked Data are:
Scenario 1: Helping clinicians to get argumentative graphs to analyze clinical trials and make evidence-based decisions.
Scenario 2: Helping hospital physicians to collect ranges of human organism's substances (e.g., cholesterol) from scientific articles, to determine if the substances' levels of their patients are normal or not.
Scenario 3: Helping missions heads from a Cancer Institute to collect scientific articles about cancer and coronavirus to elaborate research programs to deeper study the link between cancer and coronavirus.

In this section, the pipeline developed as part of this project to create Linked Data from the CORD-19 dataset is presented. A superficial view on the pipeline structure with its various components and functions is given. Subsequently, the creation of the argumentative RDF subgraph with ACTA is detailed and an example showcased.

8.2.1 Covid-on-the-Web RDF dataset

In just a few weeks, several tools were deployed to analyze the *COVID-19 Open Research Dataset* (CORD-19) [145], that in the first versions already gathered over 50,000 full-text scientific articles related to the coronavirus family. In this context, also ACTA was applied and improved on the occasion. Besides my work on ACTA, the vast expertise in the team in the management of data extracted from knowledge graphs, both generic or specialized, allowed to enrich the CORD-19 dataset from different sources. In particular, DBpedia Spotlight [146], Entity-fishing[9] and NCBO BioPortal Annotator [147] were used to extract Named Entities from the CORD-19 articles, and disambiguate them against Linked Open Data (LOD) resources from DBpedia, Wikidata and BioPortal ontologies.

LOD is the freely available part of Linked Data, which is structured data on the Web. Linked Data is an essential part of the Semantic Web, which aims at making the Web data machine-readable to allow semantic queries and reasoning. The Semantic Web is the shift in paradigm away from representing what exists on the Web, i.e., in form of URLs, towards representing one the Web what exists. This can be entities, concepts or properties (resources), which are identified by an Uniform Resource Identifier (URI) or Internationalized Resource Identifier (IRI). The properties of resources, which characterize them, are modeled as relations, which are defined in ontologies. Each relationship of a resource is described with a triple (subject, predicate, object). For example, the triple `<dbr:Paris><dbo:country><dbr:France>` represents the fact that Paris is located in France. This triple is given in the RDF schema, which

[9] https://github.com/kermitt2/entity-fishing

is the standard model for data interchange on the Web. The *dbr* and *dbo* are prefixes that encapsulate the full URI, e.g., *dbr* is short for `http://dbpedia.org/resource/`. The standard query language for RDF is SPARQL. (SPARQL) endpoints serve as interfaces to query the data. Since the goal of the Covid-on-the-Web project was to create Linked Data about the coronavirus, the information gathered by tools, like ACTA, needed to be converted. To this end, the Morph-xR2RML[10] platform, which is a tool that maps relational or non relational databases to RDF, turned the result of the mining tools into the *Covid-on-the-Web RDF dataset* and a public SPARQL endpoint [11] was deployed to serve it. Particular attention was paid to comply with the open and reproducible science goals, and the FAIR principles[12] [148]. This openness of the data and code will allow contributors to advance the current state of knowledge on this disease which is impacting the worldwide society.

For the manipulation of the knowledge graph and the visualization and exploration, the Corese[13] [149] and MGExplorer [150] platforms were integrated. These visualization techniques are meant to help users understand the relationships available in the results. Specifically, the MGExplorer and the enclosed notebooks, which transform query results into other data structures like Dataframes, bring the potential of these technologies to other fields, e.g., the biomedical and medical ones. All these tools are fused into one integration pipeline [144], as depicted in Figure 8.6. The genericity of the basic tools allow the later application of the resources to a wider set of scenarios. In the current state, this pipeline facilitates the extraction

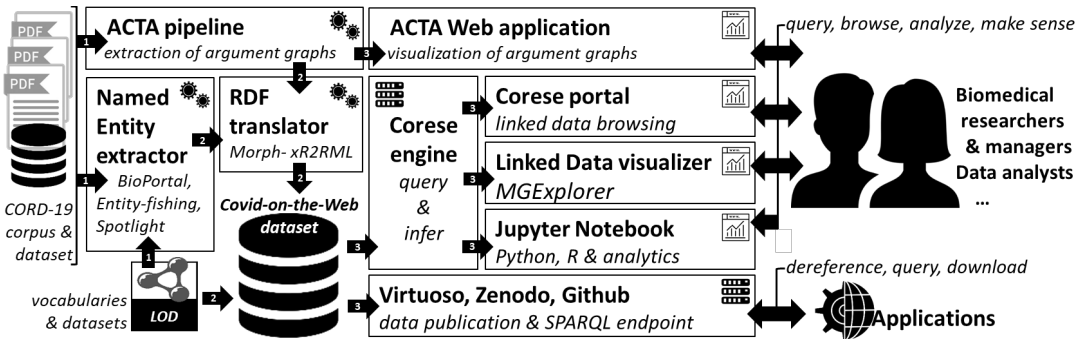

FIGURE 8.6: Illustration of the Covid-on-the-Web [144] pipeline, its services and applications.

and visualization of information from the CORD-19 dataset through the production and publication of a continuously enriched LOD knowledge graph. It is intended to engage in a sustainability plan aiming to routinely ingest new data and monitor knowledge base evolution so as to reuse updated models. For example, before the pandemic the SARS-Cov-2 entity was not existing in Wikidata. Moreover, since the emergence of the COVID-19, the unusual pace at which new research has been published and knowledge bases have evolved raises critical challenges. Therefore, the

[10] https://github.com/frmichel/morph-xr2rml/
[11] https://covidontheweb.inria.fr/sparql
[12] Fair: **f**indability, **a**ccessibility, **i**nteroperability, and **r**eusability
[13] https://project.inria.fr/corese/

knowledge graph is updated with newer releases of the CORD-19 dataset, and the extraction and disambiguation models to stay current with the latest development.

Overall, the increasing COVID-19 literature is of interest for health organisations and institutions to extract and intelligently analyse information on a disease which is still relatively unknown and for which research is constantly evolving. This necessarily leads to debates and numerous controversies regarding the origin, diagnosis and treatment of the disease [151]. What researchers need are tools to help them get convinced that some hypotheses, treatments or explanations are indeed relevant or effective, etc. Exploiting argumentative structures while reasoning on named entities can help address these user's needs and so reduce the number of controversies or at least offer the possibility to get informed. This combination of argumentative structure, PICO elements and named entites, is a unique feature with respect to other aggregated COVID-19 related datasets, such as **CORD-19-on-FHIR**[14], **KG-COVID-19**[15] or **CKG-COVID-19**[16], which all focus on biomedical ontologies.

8.2.2 CORD-19 Argumentative Knowledge Graph

As can be seen in Figure 8.6, the Covid-on-the-Web RDF dataset consists of two major knowledge graphs: the *CORD-19 Named Entities Knowledge Graph* (CORD19-NEKG) and the *CORD-19 Argumentative Knowledge Graph* (CORD19-AKG). The elaborated creation of the former will be skipped, since it is not related to the main topic of this thesis. For more details, I refer the interested reader to our paper published at the International Semantic Web Conference 2020 [144]. The main focus of this section will be the re-purposing of ACTA to annotate the CORD-19 dataset resulting in the CORD19-AKG. In general, the base functions of ACTA are the same as described in Section 8.1. It retrieves the main claim(s) stated in the trial, as well as the evidence linked to this claim, and the PICO elements. The first notable difference is the extension of the link prediction to include proper relation types. Other differences comprise the change of the pre-trained transformer weights and the alignment and linking of the output to ontologies.

Practically, each abstract of the CORD-19 dataset was analyzed by ACTA and translated into RDF to yield the *CORD-19 Argumentative Knowledge Graph*. The pipeline consists of four steps: *(i)* the detection of argumentative components, i.e. claims and evidence, *(ii)* the prediction of relations holding between these components, *(iii)* the extraction of PICO elements, and *(iv)* the production of the RDF representation of the arguments and PICO elements.

Component Detection Corresponding with Section 8.1, this is a sequence tagging problem, where the adapted transformer with the RNN/CRF layer is employed.

[14] https://github.com/fhircat/CORD-19-on-FHIR
[15] https://github.com/Knowledge-Graph-Hub/kg-covid-19/
[16] https://github.com/usc-isi-i2/CKG-COVID-19

Contrary to the original version, here, the weights in BERT are initialized with specialised weights from SciBERT [63], which is pre-trained on full papers from Semantic Scholar and biomedical data, and provides an improved representation of the language used in scientific documents such as in CORD-19. Alas, since there is no reference CORD-19 subset that has been manually annotated and could serve as ground truth, it is hardly possible to evaluate the quality of the models used to extract argumentative structures on the CORD-19 dataset. Thus, the performance on the AbstRCT dataset is reported as displayed in Chapter 4 Table 4.3, where SciBERT delivers a .87 macro F_1-score on the neoplasm test set. As a final step, the components are extracted from the predicted label sequences.

Relation Classification As previously mentioned, this is one of the major changes to the ACTA pipeline. The multiple choice model for link prediction was replaced by a model considering the types of relations. As defined in Chapter 3, two types of relations can hold between argumentative components. The *attack* relation holds when one component is contradicting the proposition of the target component, or undercutting its implicit assumption of significance, i.e., stating that the observed effects are not statistically significant. The *support* relation holds for all statements or observations justifying the proposition of the target component (even if they justify only parts of the target component). Determining which relations hold between the components is treated as a three-class sequence classification problem, where the sequence consists of a pair of components, and the task is to learn the relation between them, i.e. *support*, *attack* or *no relation*. This corresponds to the SentClf approach explained in Chapter 4.2. As for the sequence tagging task, the SciBERT transformer is used to create the numerical representation of the input text, and combined with a linear layer to classify the relation. The model is fine-tuned on the AbstRCT dataset for argumentative relations resulting in .68 f_1-score on the neoplasm test set (compare with Chapter 4 Table 4.5).

PICO Element Detection The same architecture as for the component detection is employed. The model is still trained on the EBM-NLP dataset [6] to jointly predict the participant, intervention and outcome candidates for a given input. Also here the SciBERT pre-trained weights are used. Contrary to the base version of ACTA, where the whole abstract is annotated, for the application on the CORD-19 dataset, only the argumentative components are annotated with the PICO elements they contain. This change was done, since the Argumentative Knowledge Graph is supposed to focus on the contained arguments, while the Named Entities Knowledge Graph holds information about the (full text) document. As mentioned in Section 8.1, PICO elements can occur multiple times in various forms through out the trial abstract. By annotating each argument component with its contained PICO elements, a redundant overall list as it was the case in the base version is avoided. However, the problem of aligning the various appearance of the elements still remains. Since this

is not a straightforward task, the most practical approach seemed to link each PICO element to concepts in the Unified Medical Language System (UMLS), which is a controlled vocabulary for a standardized communication of biomedical concepts. This way, it does not solve redundancy detection, but facilitates structured queries, which is after all a fundamental keystone of the Covid-on-the-Web RDF dataset. The linking is done via the ScispaCy [152] entity linker. Annotation bodies are the UMLS concept identifiers (CUI) and semantic type identifiers (TUI). These outcome elements have been further enriched by the Effect-on-Outcome as described in Chapter 6. With this, the queries can be even more advanced.

RDF Representation To represent the extracted information in a standardized way for interchange on the Web, proper ontologies had to be selected. In particular, the *CORD-19 Argumentative Knowledge Graph* draws on the Argument Model Ontology (AMO)[17], the SIOC Argumentation Module (SIOCA)[18] and the Argument Interchange Format[19]. AMO[20] is an ontology describing arguments following Toulmin's model of argument [86]. Each argument identified by ACTA is modelled as an `amo:Argument` to which argumentative components (claims and evidence) are connected. The claims and evidence are themselves connected by support or attack relations (`sioca:supports/amo:proves` and `sioca:challenges` properties respectively). The SIOCA module was used complementary to cover for the attack relations, which are not defined in AMO the way they are needed to describe the output of ACTA. Finally, AIF was used as a third ontology to include the textual description of the argument components and achieve compatibility with the Argument Web [153]. Following these ontologies, Listing 8.1 sketches a shortened example of one argument (lines 7-11) and one evidence (lines 13-18). The first line of the resource is the URI, for the document this is line 7 and for the evidence line 13. In line 8 one can see that the resource in line 7 is an `amo:Argument` occurring in the resource (document), which URI is stated in line 9. The argument contains an evidence (line 10), where the last digit is the component ID (0), and a claim (line 11). For reasons of clarity the remaining evidence 1-5 are not listed. However, evidence 0 is presented exemplary. Line 14 specifies the type in all three ontologies, i.e., that it is a `amo:Evidence`, `sioca:Justification` and `aif:I-node`. Line 15 refers to the resource the evidence occurs in. Line 16 holds the actual text of the evidence, i.e., what was extracted by ACTA. The last two lines describe the relation of the evidence to the claim, which in this case is a *support* relation modeled as `sioca:supports` and `amo:proves`. Additionally to the argumentative information, the PICO elements are added to the annotation bodies with their UMLS CUI and TUI, as described above.

[17] `http://purl.org/spar/amo/`
[18] `http://rdfs.org/sioc/argument#`
[19] `http://www.arg.dundee.ac.uk/aif#`
[20] `https://sparontologies.github.io/amo/current/amo.html`

```
@prefix prov:   <http://www.w3.org/ns/prov#>.
@prefix schema:  <http://schema.org/>.
@prefix aif:    <http://www.arg.dundee.ac.uk/aif#>.
@prefix amo:    <http://purl.org/spar/amo/>.
@prefix sioca:  <http://rdfs.org/sioc/argument#>.

<http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496...>
  a               amo:Argument;
  schema:about    <http://ns.inria.fr/covid19/4f8d24c531d2c33496...>;
  amo:hasEvidence <http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496.../0>;
  amo:hasClaim    <http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496.../6>.

<http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496.../0>
  a amo:Evidence, sioca:Justification, aif:I-node;
  prov:wasQuotedFrom <http://ns.inria.fr/covid19/4f8d24c531d2c33496...>;
  aif:formDescription "17 patients discharged in recovered condition...";
  sioca:supports  <http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496.../6>;
  amo:proves      <http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496.../6>.
```

LISTING 8.1: Example representation of argumentative components and their relation in RDF.

Data Extraction Since the CORD-19 is a mainly automatically created dataset, it contains errors, like cut off or empty abstracts. Hence, regarding the extraction of the data, only the abstracts longer than ten sub-word tokens were processed by ACTA to ensure meaningful results. Inputs were tokenized with the BERT tokenizer, where one sub-word token has a length of one to three characters. In total, almost 30,000 documents matched this criteria in CORD-19 v7, on which the first complete published dataset was based, and 68,000 in CORD-19 v47, which is the latest version the pipeline was applied on at the time of writing. ACTA was deployed on a 2.8GHz dual-Xeon node with 96GB RAM. The data was split into batches of 5,000 documents, processing each batch took approximately 3 hours. The output JSON documents were loaded into MongoDB and translated to the RDF model using Morph-xR2RML. The translation to RDF was carried out on the same machine as above, and took approximately 10 minutes. In compliance with the open science principles, all the scripts, configuration and mapping files involved in the pipeline are provided in the project's Github repository [21] under the terms of the Apache License 2.0, so that anyone may rerun the whole processing pipeline (from articles mining to loading RDF files into Virtuoso OS).

To sum up, this section described the data and software resources provided by the Covid-on-the-Web project with an extra focus on the contribution of the ACTA system developed in the context of this thesis. Various tools to process, analyze and enrich the CORD-19 dataset, were adapted and combined to make it easier for biomedical researchers to access, query and make sense of COVID-19 related literature. The output of the pipeline is published as a Linked Data knowledge graph describing the named entities mentioned in the CORD-19 articles and the argumentative graphs they include. On top of this knowledge graph, other members of the project adapted and deployed several tools providing Linked Data visualizations, exploration methods and notebooks for data scientists. Through active interactions

[21] https://github.com/Wimmics/CovidOnTheWeb/

(interviews, observations, user tests) with institutes in healthcare and medical research, it is ensured that the Covid-on-the-Web pipeline is guided by and aligned with the actual needs of the biomedical community. This proves the (re-)usability of the work presented in this thesis as a means to assist in documentary research and thus, in the deliberation process of clinicians, which was the initial goal.

Chapter 9

Related Work

This chapter presents and discusses the related work in the context of evidence extraction and argumentation-based applications in Evidence-Based Medicine. It further sets the work presented in this thesis into perspective by showcasing the related development in the AM field and pointing out differences to existing approaches.

This section highlights various approaches to (semi-)automatically assist in the clinical decision making process or other procedures in EBM. In particular, Section 9.1.1 introduces various argumentation-based approaches, which have been employed to evaluate medical evidence or provide decision support. Section 9.1.2, presents approaches targeting automated evidence extraction and its precursors. Subsequently in Section 9.2, an overview over recent developments in the relevant Argument Mining fields, i.e., component detection, relation classification and evidence type classification, is provided.

9.1 Applications in Evidence-based Medicine

9.1.1 Argumentation-based Decision Support

Argumentation-based decision making is becoming increasingly prominent in healthcare applications. Several formal frameworks have been proposed to tackle the issues of reasoning upon clinical evidence and detecting possible conflicts in medical knowledge bases [8–10, 19, 154]. Different kinds of data can be explored in this context, e.g., clinical trials, clinical guidelines, electronic health records, combined with the patient and clinician preferences and the specific constraints raised by the particular medical branch taken into account. The general aim of such approaches is to support clinicians and practitioners in taking informed decisions. However, the main limitation of these approaches is that they assume the availability of structured information, e.g., in the form of databases or knowledge bases.

For example, Longo and Hederman [10] investigate defeasible reasoning on the breast cancer recurrence prediction. A knowledge base of 277 breast cancer patients

serves as a basis to construct structured defeasible arguments, which are then evaluated for argument justification status using acceptability semantics. The knowledge base consists of attributes about the patient, such as age, tumor size, applied irradiation and recurrence of cancer. They argue that clinicians prefer a decision-making support system, which delivers explanations rather than just numerical values and that property of explainability is provided with argument-based frameworks.

Another approach to find contradicting conclusions, Qassas et al. [19] propose an argumentation scheme based approach to analyse clinical discussions. The motivation is to help discover weak points, such as missing evidence, invalid reasoning and hidden assumptions in the debate about the choice of treatment. As mentioned in Chapter 2.2, argumentation schemes can be target of the argumentative structure prediction. Highlighting conflicting diagnostic hypotheses applying these schemes is an example of a meaningful use case, which is not the case for analysing single clinical trials.

Moving to the argumentation approaches on clinical trials, Craven et al. [9] apply assumption-based argumentation (ABA) [155] to clinical trials of breast cancer drugs. To this end, statements in the trial are translated into logical sentences and bigger arguments. An OWL ontology for the medical domain is created and ground rules for the argumentation are derived from it. While clinical trials do indeed serve as a fruitful source to create these logical sentences, which allow the automation of the decision making process, the creation of the proper data structure is labour-intensive and done manually so far. Same applies for the creation of the ontology. In the paper only 57 papers were annotated giving an idea of the amount of work which is necessary for the translation into logical sentences. A different aspect of clinical trials is targeted by Hunter and Williams [8, 154]. They propose a framework to represent and synthesize knowledge from clinical trials, which they call evidence aggregation. The knowledge comes in the form of inductive and meta arguments, where inductive arguments are propositions that one treatment is superior, equivalent or inferior to another one with respect to a certain indicator/outcome given an evidence. Meta arguments are counterarguments weakening the inductive arguments, e.g., that an evidence was not statistically significant or the trial setup contains flaws. From this, an argument graph is constructed. Based on the preference criteria, such as outcomes and their magnitude, and evidence quality, the argument graph is evaluated to determine which treatment is superior. With respect to the automatic processing pipeline of trials described in this thesis, the final data structure provided is relatively close to what is required for this argumentation framework. Claims and evidence with the added information about Effect-on-Outcome fulfill the requirements as inductive arguments, while some of the evidence (types), i.e., evidence stating that an effect is statistically not significant, are meta arguments. Together with other information about the trial setup (possible meta arguments – see

the next section for examples of automatic extraction) most of the necessary information for this argumentation approach can be extracted automatically from unstructured text. This further demonstrates the relevance of the work presented in this thesis, complementing the application scenarios discussed in Chapter 8.

9.1.2 Automated Analysis of Clinical Trials

Besides the aforementioned argument-based approaches, various other works exist analysing and processing the information of clinical trials. Most of them do not have the goal of providing structured data for argumentation frameworks, but rather offer a way of automatically extracting or distilling knowledge about the trials in human readable form. Therefore, only few of them produce structured data which can be reused in argumentation frameworks.

Comparisons are a crucial part of the scientific exchange and communication. Usually, the newly proposed method is compared with an established method to demonstrate the benefits and/or superiority. As described in Section 2.1, in the medical domain comparative studies are the prevalent type of considered studies, where researchers evaluate the effectiveness, risks and side-effects of a drug compared with a control intervention. Accordingly, reports of clinical trials are rich on comparative statements. Therefore, early work focused on the detection and extraction of these comparative structures. Park and Blake [115] propose an approach to automatically detect comparison claims in full-text scientific articles, achieving a .74 F_1-score. They experiment with semantic and syntactic features in Naive Bayes, SVM and Bayesian networks. The lexical features comprise handcrafted dictionaries capturing characteristic of comparative sentences. For example, if a sentence contains any directed verbs, inflections marking a comparisons, like *"better"*, or other cues, like *"above"* or *twice of*. While handcrafted features used to be a common component in early Machine Learning in NLP, I decided against developing own features, because at least for lexical information, most of the characteristics should be covered by the bag-of-words model, which was used for the first experiments. With neural networks and their automatic feature extraction from word embeddings or with the attention-based transformers at latest, lexical features became obsolete. As syntactic features, the authors developed rules for the dependency graphs, which is intuitive since in the case of comparative structures syntax can be a distinctive characteristic. This observation, together with the subsequently described approach, which also utilizes syntactic information, further affirmed the choice of syntax-based Tree Kernels for the experiments on argument component detection, as described in Chapter 4.1.1. As anticipated, Gupta et al. [116] do pattern matching on the output of syntactic structure and dependency parsers to extract comparison structures from biomedical texts. They work on gene expression studies, similar to the AbstRCT dataset they work only on the abstracts of RCTs, and achieved a F_1-score of .87 for comparison sentence identification. Contrary to the previous approach, the pattern matching allows also the identification of the entities involved in the comparison

and their aspects. The deeper analysis of a comparison, such as entities and aspects, are an important step towards the automatic information extraction and formalization of clinical trials. Earlier to Gupta et al. [116], Fiszman et al. [156] tackled this task with under-specified semantic interpretation. The goal was to identify the entities in comparative sentences and express one entity A in terms of the other entity B, e.g., A is superior to B. Based on handcrafted linguistic patterns, the comparative constructions are processed and automatically analysed. The authors state that approximately 30-40% of drugs and comparative statements are not recognized this way. Compared to the Argument Mining task on clinical trials, comparative structures are important, but only cover parts of the desired information. Moreover, comparisons can occur as either being a conclusive statement or an observation of experimental outcomes. Without the argumentative distinction related to their credibility, i.e., the classification into evidence and claims, decisive information relevant for the overall decision making is missing. However, the aspect of comparison with respect to a certain outcome is an imperative part of EBM practice. Thus, the addition of the Effect-on-Outcome analysis to the Argument Mining pipeline is of utter importance. Driven by the same motivation, recently, Lehman et al. [89] proposed an approach to infer if a study provides evidence with respect to a given intervention, comparison intervention and outcome. Additionally to the classification of the whole document, the model returns a sentence from the document supporting the classification result. These *rationals* [157], which are mostly comparative sentences, are important evidence which support the classification result in a human readable way. This approach is similar to the one presented in this thesis in the sense that the overall goal is to determine the effect an intervention has on an outcome and provide evidence for it. Contrary to the here presented approach, they focus more on finding one rational, which is called an argument component in the AM context, objective oriented for the given prompt. A prompt consists of three PICO elements, i.e., the intervention, the comparison intervention and one specific outcome. In the context of EBM a prompt can be considered as a well-built clinical questions [25]. The authors experiment with neural networks, in particular with GRUs and attention mechanisms. Two possible architectures are proposed. Firstly, a pipeline approach, where the evidence is detected in a first step. Contrary to argument component detection presented in Chapter 4.1.2, they classify complete sentences without segmenting them further, which is an essential part in Argument Mining [33]. Subsequently, the found evidence are used to classify the document with respect to the given prompt, i.e., a document states that the outcome in the prompt was either *significantly increased/decreased* by the intervention with respect to the comparison intervention or that there was *no significant difference*[1]. The second proposed approach is an end-to-end architecture trying to learn both information jointly, where they achieve a F_1-score of .52 for the document classification. The F_1-score is relatively low, because they conduct

[1]For the reasons mentioned in Section 3.2.3, I deviate from this annotation scheme by having a wider spectrum of outcome status.

a document wide target oriented search for a specific evidence, which is an (unnecessarily) challenging task. The AM pipeline is setup in the opposite direction, where the first step is to find evidence in the form of an argumentation graph, which is human readable and then, in a subsequent step, enrich these graphs with information about the contained PICO elements, e.g., if an outcome was increased. This data-driven approach has multiple advantages over the work proposed in [89]. First, it requires less computational capacities, because the pipeline is run once for each trial and not x-times for every single prompt (usually one wants to have information about more than one outcome). Secondly, additionally to the outcome description, the argumentation graph contains outcome unspecific information, which is relevant in judging the results of a study. For example, limitations of the study where the authors state that their findings need further confirmation.

As explained in Chapter 2.1, in search for relevant evidence, practitioners of EBM use a specialised framework called PICO, which stands for ***P**atient Problem* or *Population*, ***I**ntervention*, ***C**omparison* or *Control*, and ***O**utcome*. Searching for relevant trials and finding meaningful answers is a time consuming and laborious task for clinicians. Automating this process of evidence collection from documents could unburden the clinicians substantially. Besides the work proposed in this thesis, there have already been other systems assisting in automatic evidence extraction. Contrary to my work, they focus more on assisting in the semi-automatic completion of evidence tables instead of interpreting the results. One early example for this is ExaCT [158]. The system extracts information containing PICO elements based on a SVM. It was designed to search full text articles, but was limited by the scarce training data available. Whereas nowadays, there is the EBM-NLP dataset [6], which is a collection of considerable size of sentences annotated with PICO elements. Similarly, Jin and Szolovits [7] propose a NN based on word2vec embeddings, a LSTM and a CRF to classify sentences as belonging to a certain PICO category. Regarding finer token-level classification, Trenta et al. [5] proposed a maximum entropy classifier to mine characteristics of randomized clinical trials in form of PICO elements. Their corpus comprises 99 manually annotated abstracts, which are used as a basis to start the AbstRCT data collection. The importance of PICO elements is undeniable and strongly motivates the aforementioned work. That is why the PICO element detection module is integrated in ACTA, which was made possible by the release of the sizeable EBM-NLP dataset.

Another system facilitating the evidence gathering process is RobotReviewer [159, 160], which summarizes the key information of a clinical trial. These key information comprise the interventions, trial participants and risk of bias, where the latter is related to finding potential design flaws of the studies. Flaws in the study design can mean to check if participant groups were randomly allocated, or if the study was double-blinded. So neither participants nor medical personnel knew who was given the intervention or comparison intervention. The tool is open source[2] and was

[2] Available here: `https://github.com/ijmarshall/robotreviewer`.

recently updated to use BERT embeddings (SciBERT) to keep up with the current development in NLP.

With respect to AM, one of the few studies in focusing on the biomedical domain was presented by Green [16, 18, 161], who proposed argumentation schemes and inter-argument relations for the annotation of arguments in research articles. Yet, such annotation schemes are only partially applicable for argument extraction in RCT abstracts and the proposed work is purely of theoretical nature. Accordingly, before AbstRCT no huge annotated dataset for AM was available for the healthcare domain. There exists a dataset of contradicting claims [162], which was created using research abstracts of studies considered in systematic reviews related to cardiovascular diseases, but this dataset does not contain the corresponding evidence backing those claims. Furthermore, Blake [163] proposes a claim framework for biomedical literature to reflect how the authors communicate their findings. Claims are classified into explicit, implicit, correlation, comparison and observation. While this is an interesting facet for evaluating the qualitative side, for the mining approach proposed in this thesis this distinctions are unnecessary, since their function in the argumentative structure does not differ. However, it is a promising direction for future research addressing the evaluation of an argument's strengths in this domain.

9.2 Argument Mining

As stated in Chapter 2.2, Argument Mining comprises various subtasks by now. Due to the increasing size and diversity of tasks, this section limits its overview of the related work to the subtasks which are relevant for the direct context of this thesis. These are two standard tasks – *(i)* the identification of arguments within the text and the identification of their textual boundaries, and *(ii)* the prediction of the relations holding between the arguments identified in the first stage – and evidence type classification. For these tasks different methods have been employed, ranging from Support Vector Machines over Naïve Bayes classifiers to Neural Networks.

While AM methods have been applied to heterogeneous types of textual documents, e.g., persuasive essays [85], scientific articles [28], Wikipedia articles [164], political speeches and debates [165], and peer reviews [166], most of the approaches consider only single aspects of the Argument Mining pipeline. Few approaches consider the whole AM pipeline in different application scenarios. In particular, Stab and Gurevych [85] propose a feature-based Integer Linear Programming approach to jointly model argument component types and argumentative relations in persuasive essays. Differently from the AbstRCT data, essays have exactly one major claim each. The authors impose the constraint such that each claim has no more than one parent, while no constraint holds for the methods presented in this thesis. Due to the independent pairwise classification, this can lead to an undesirable divergent argument structure in the case of the relation classification using the SentClf method, as described in Chapter 4.2. In contrast with this approach, Eger et al. [36]

present neural end-to-end learning methods in AM, which do not require the handcrafting of features or constraints, using the persuasive essays dataset. They employ a TreeLSTM on dependency trees [106] to identify both components and relations between them. The core idea of TreeLSTMs is to leverage the recurrent nature of the LSTM on the tree structure. The bidirectional structure of the TreeLSTM propagates information from the leaves to the root node and vice-versa. In particular, the employed TreeLSTM from Miwa and Bansal, which Eger et al. evaluated for AM, consists of three layers. A shared embeddings layer, a word sequence layer and a dependency tree layer. The sequence layer consisting of a BiLSTM which is responsible for the token-wise entity/argument component detection, while the TreeLSTM layer is responsible for the relation classification. Both layers share the same word embedding layer and the sequence layer further forwards the token-wise hidden state output of the BiLSTM to each node in the dependency tree. The key point is that the tree-structured LSTM-RNN (TreeLSTM) allows shared weight matrices for same-type children. More details about this approach can be found in [106]. For the application of this TreeLSTM to AM, they decouple component and relation classification labels, which are jointly learned, using a dependency parser to calculate the features. In the pipeline proposed in this thesis, not only the label spaces but also the two classification tasks are decoupled, in line with the claim in [36] that decoupling component and relation classification improves the performance. Furthermore, the same work addresses component detection as a multi-class sequence tagging problem [167]. Differently from their approach, which does not scale with long texts as it relies on dependency tree distance, the approach proposed in this thesis is using distance independent attention mechanisms to counter this problem. In addition, whilst persuasive essay components are usually linked to components close by in the text, in the AbstRCT dataset links may span across the whole RCT abstract. The joint learning is achieved by sharing weights among the different tasks, and the sequence tagging is done independently of the context. Finally, in [36], each word is a feature, while most of the methods evaluated in this thesis opt for sub-word level.

Ajjour et al. [168] proposed a deep learning approach for segmentation of text into argument units. Here, the task is, again, formulated as a sequence tagging problem, where a label is assigned to each token following the BIO-tagging scheme. The authors only tackle the argument unit segmentation (argumentative vs non-argumentative) without the further classification of the components. Contrary to the performed five class argument component detection, this translates to a three class classification problem, i.e., *Arg-B*, *Arg-I* and *Arg-O*. The best performing model consists of two BiLSTM, where one is using word embeddings and the other syntactic, structural and pragmatic input features (one-hot vectors). Both BiLSTM outputs are concatenated and put through a dense layer before it is passed to another (upper) BiLSTM. The output of the last (upper) BiLSTM is used in the final classification

layer. The authors noted an decreased number of invalid BI sequences with the addition of the second (upper) BiLSTM. In later work, Spliethover et al. [169] further investigated this architecture with minor changes: they used solely one BiLSTM with word embeddings as input features and tested the efficacy of the second (upper) BiLSTM. Moreover, they investigated the effects of adding various attention layers. The results did not show any major changes in performance with respect to adding the second (upper) BiLSTM. Also, the addition of attention layers did not improve the results. In line with these observations, no stacked RNNs or attention layers are added for the sequence tagging architectures evaluated for argument component detection in this thesis. The idea is to reduce the number of invalid BI sequences not with a second (upper) RNN layer, but with a CRF, as described in Chapter 4.1.2.

Recent approaches for link prediction rely on pointer networks [170] where a sequence-to-sequence model with attention takes as input argument components and returns the links between them. In these approaches, neither the boundary detection task nor the relation classification one are tackled. Another approach to link prediction relies on structured learning [107]. The authors propose a general approach employing structured multi-objective learning with residual networks, similar to approaches on structured learning on factor graphs [171]. Here, the component classification, link prediction and relation classification are learned jointly, where the boundaries of the components are assumed to be already set. The model takes the source and target component plus a distance encoding as an input and outputs the labels for the components, the binary label if a link between the components exists and the label of the relation. Architecture-wise, both components are encoded with GloVe and fed into BiLSTM layers followed by dense layers for dimensionality reduction. The residual network serves the purpose to connect neurons in distant layers and communicate representations skipping intermediate layers. The model consists of three classifiers, one for each problem. For the relation classification, this approach is considered as a baseline, because they classify all possible component combinations similar to the SentClf setup. Contextualized word embeddings did find their way into the AM community. Contemporaneous to my first experiments, Reimers et al. [38] addressed topic-dependent argument classification with contextualized word embeddings achieving .63 F_1-score. Here, the goal is to classify a sentence given a topic as either being an argument for or against the topic, or not being an argument. This problem of sentence classification is in line with the other tasks, where BERT was shown to have an outstanding performance. In Chapter 7 of this thesis, this approach is analysed for its robustness. Another approach employing BERT published after I finished my work on the relation classification is *AMPERSAND* [172]. There, the authors address the AM tasks of component classification and structure prediction in a dialogical setting. They leverage BERT for intra- and inter-turn relation classification. First, they fine-tune the pre-trained BERT model on distant-labeled data to compensate for the relatively small size of the actual target dataset. Subsequently, this fine-tuned BERT is further fine-tuned on the actual

dataset for relation classification. Additionally to the single BERT prediction, they train a XGBoost Classifier with categorical discourse relations of the components as features to predict the argumentative relations. The final relation is determined in an ensemble method with both classifiers. This is another example supporting the integration of discourse information in transformer models, which was discussed in Chapter 7.4. Another interesting approach proposed by the authors, which could be applicable in future work for the relation prediction across multiple clinical trials, is a candidate target selection to reduce the false positives created by the exhaustive combination of all argument components. With respect to the medical domain, contrary to the proposed summarization-based selection, for cross trial relations PICO elements might be a better limiting category. However, the last two discussed approaches have in common that they assume argument components as given, and boundary detection is not considered, which is different from my approach. The work presented in this thesis is the first work to create a sequence tagging model for component classification and boundary detection utilizing the power of pre-training transformer models. In line with the work of Reimers et al., first experiments considered the BERT_{base} [59] model to address parts of the AM pipeline. Further, contrary to this preliminary work and the above mentioned related work, various contextualized Language Models and architectures are employed and evaluated on each task to span the full AM pipeline as well as the outcome analysis. In later work, Niven and Kao [129] apply BERT to the Argument Reasoning Comprehension Task (ARCT) [113], but found that the high performance of BERT is due to unevenly distributed linguistic cues. The problem ARCT poses is that given a claim, a reason and two warrants, the correct warrant needs to be selected. To solve this task a deep comprehension of the presented argument is necessary, which the authors deny that BERT has learned it.

Concerning the evidence classification little work has been done. Rinott et al. [97] tackled this problem on Wikipedia based data, dividing the evidence into *study, anecdotal and expert* evidence. This taxonomy is not applicable to all types of data, i.e., clinical trials, since all of the documents are studies. For this reason, a more fine-grained taxonomy adapted for clinical trials was developed, as described in Chapter 5. With regards to the general structure of scientific publications, there has been work analysing it from the perspective of discourse structure [173, 174]. Based on this, Kirschner et al. [175] proposed a combined annotation scheme where the two argumentative relations (support and attack) are complemented with two discourse relations from Rhetorical Structure Theory [176] (sequence and detail). The *detail* relation is used, if a component gives more information about another component without argumentative reasoning, for instance in the case of definitions. The authors annotated full-text articles from educational research. In this context definitions occur more often and the addition of this relation can be justified, but in the context of RCT abstracts this relation becomes obsolete, since definitions are not given in an abstract. Further, as Green [161] observed, many arguments in the biomedical

domain have implicit warrants. Thus, even for a full-text analysis of clinical trials, the benefit of this relation needs to be evaluated. Concerning the *sequence* relation, this is used when argument components belong together and require each other, for instance to support a conclusion. This corresponds to the linked argument structure mentioned in Chapter 3. Since pieces of evidence that fall under this category in the AbstRCT dataset occurred next to each, it was decided to annotated them as one argument component instead of introducing the *sequence* relation.

Chapter 10

Conclusion and Future Perspectives

Clinical decision making is often intricate to a high degree. Especially with the growing number of published medical studies on the Web, the selection, assessment and application of these trials as relevant evidence for the decision making becomes a challenging and laborious task. This sets the need for systems to (semi-)automatically assist in processing this huge amount of data. While there are decision support systems, they require structured data, extending the demand from processing the data into a human-readable format to extracting and preparing the data into a machine-readable format. The work in this thesis addressed these two major problems: (i) the problem of researching clinical evidence; for this, an automated approach was proposed to supply clinicians with valuable information about clinical trials. In particular, this thesis presented an Argument Mining approach for processing and analysing clinical trials with respect to their argument components, such as claims and evidence, and the relations (attack or support) between them. (ii) the demand of structured data for the aforementioned decision support systems; additionally to the acquired argumentative information, the results of a trial are further put into a machine-readable format by additionally aggregating the argumentative structure with further medical information in form of PICO elements and the Effect-on-Outcome analysis.

In particular, to provide these solutions, the research questions introduced in Chapter 1 were addressed resulting in the following contributions:

1. Creation of the AbstRCT Dataset The dataset was created from a collection of RCT abstracts from the MEDLINE database via PubMed. A bipolar argumentation scheme for argument components, such as claims and evidence, and their relations, i.e., support and attack, was applied to annotate the collected data. The conducted annotation study is portrayed in Chapter 3. Besides the annotated argumentative information, the Effect-on-Outcome scheme was developed and the dataset subsequently annotated with it. The latter scheme focuses on the medical information content of the argument components, for example, it encodes that an intervention increased or decreased a certain outcome. The data annotation resulted in a Fleiss'

kappa of 0.68 for argument components and 0.62 Fleiss' kappa for the relations. The effects on the outcomes associated to the identified argumentative components showed a Fleiss' kappa of 0.81, all representing at least substantial agreement and thus attesting the reliability of the created dataset. In total, the second version of the AbstRCT dataset contains 660 abstracts spanning the topics of neoplasm (500 abstracts), glaucoma (100 abstracts), hepatitis b, diabetes and hypertension (20 abstracts each).

2. Domain-specific Argument Mining and Outcome Analysis Pipeline The thesis introduced a full Argument Mining pipeline with an outcome analysis extension. The development described in Chapter 4 started with first experiments with feature-based SVMs on Tree Kernels to classify sentences into either argumentative, claim or evidence. The evidence are further subdivided into the more fine-grained labels *comparative*, *significance*, *side-effect* and *other* to provide a better structure with regards to the reasoning process in argumentation-based systems or filtering/querying for specific information. Subsequently, a full AM pipeline was developed considering both, the detection of argument components and their boundaries, and the argumentative structure prediction. For the former, a sequence tagging approach was employed combining a domain specific BERT model with a GRU and CRF on top to identify and classify argument components. The relation classification task was cast as multiple choice problem and was compared with recent transformer models for sequence classification. The proposed approach significantly outperformed standard baselines and previous state-of-the-art AM systems with an overall macro F1-score of .87 for component detection and .68 for relation prediction. The Effect-on-Outcome analysis introduced in Chapter 6 presents a major extension of the aforementioned AM pipeline. There, outcomes mentioned in the argumentative components were detected and their effect subsequently classified, i.e., if an intervention has *Improved*, *Increased* or *Decreased* the outcome, or that there was *NoDifference*, or *NoOccurrence* of the outcome. The experiments relied on the second annotation scheme in the AbstRCT dataset, mentioned above. The introduced pipeline for this analysis consists of two parts. First, an outcome detection module and second, an effect classifier. The former employed a sequence tagging approach to detect mentioned outcomes. For this, the same sequence tagging architecture as for the argument component detection was used. The effect classifier subsequently classifies the extracted outcomes with respect to the aforementioned labels. This task was cast as sequence classification and thus, the same architectures as for the relation classification were employed. The pipeline achieves a macro F1-score of .80 for Effect-on-Outcome classification.

3. Proof-of-Concept and Limitation Analysis In an extensive evaluation the errors of the system were analysed highlighting the shortcomings of the employed

architectures and methods, especially with regards to the classification of argumentative relations. Additionally, an investigation was undertaken to analyse the general robustness of the underlying bidirectional transformer model in Chapter 7. To this end, different ways to produce meaningful adversarial examples were investigated. The quality of the generated perturbations was assessed in a user study and the effect of adversarial training for argument classification was empirically evaluated. The obtained results attest a relatively reliable handling of input with simple linguistic variations (in 5-10% of the cases the perturbation of the input changed also the prediction of the model). Furthermore, general weak points of the transformer model were highlighted to indicate that this solution is still imperfect. However, the applicability of the work proposed in this thesis was demonstrated with a Proof-of-Concept system (ACTA) in Chapter 8, which illustrated the impact of the argumentative information in interplay with the PICO elements. For instance, this hybrid system can identify when a claim reports an outcome as being *safe* or *efficient*, but also that the associated side effects are classified as *increased*, setting the claim into perspective. This combined analysis reveals more fine-grained categorization of the statements in RCTs. Its re-usability was further shown in the context of the Covid-on-the-Web project, where it was adapted and integrated in a pipeline creating Linked Data. In the resulting argumentative knowledge graph both, the PICO elements and the argument components, are represented with ontologies to enable semantic queries. This is an important step towards machine-readable data to support an automated analysis of trials. While an automated analysis of trials seems like it could harbour the risk that decisions concerning a patient's treatment might be automated, this was not the final objective. The methods proposed in my thesis particularly serve the purpose to provide assistance for a clinician to take informed decisions with respect to the current development in research. The final choice of treatment has to be made by the clinician after a case specific evaluation of the presented information, together with the patient's preferences. The extracted arguments for or against certain treatment options do not only help in taking this decision, but also in explaining the decision in a reasonable way to the patient. For the latter, especially a visualization, as for instance provided by the ACTA system, can bring the decision closer to the patient. After all, one goal of Argument Mining is to make cumulative information more accessible, which was ultimately shown to be achieved by the work presented in this thesis.

In brief, the research conducted in the context of this thesis showed how to employ and develop Argument Mining methods for the medical domain, in particular clinical trials. As the field is still evolving, and to foster future research in the area of Argument Mining on clinical trials, also with respect to the (medical) analysis of observed outcomes, the AbstRCT dataset was made available together with the source code of the experiments. I believe the above listed contributions of my work are valuable input to motivate the community to build upon this work and spur the reuse and adaptation of the dataset.

Future Perspectives

While important concepts have been carved out in my work, it leaves space for further research directions and future improvements.

First, concerning the choice of model architectures, throughout this thesis several weaknesses of the employed Language Model based transformer models have been discovered showing that the current pre-trained models achieve impressive results, but are not the final answer. Especially for sequence/relation classification, there is potential to further enhance the classifier. With the employed *SentClf* approach there is a high rate of false negatives (i.e., relations that have been wrongly predicted as NoRelation). Also, the multiple choice approach could not solve the problem satisfactorily. The classification of argumentative relations requires a deeper understanding and encoding of both, the source and the target, component into a reusable representation in a semantic space. There, the models need to learn to infer the causal or consequent relationship between two components based on their representations. Furthermore, the learned relation should be abstract enough to be able to be transferred and leveraged on similar cases in new data. This is where current transformer models reach their limits. As it has been contemplated by Bender and Koller [135], shown for argument comprehension in [129] and discussed in the Chapters 4.2 and 7, the suspicion is that the model rather learns linguistic patterns or cues than a real semantic understanding. A model could predict a relation correctly, but it cannot explain the reasoning why the two linked components are in a relationship, since the required warrants are not explicitly mentioned in the text. This classification on signals purely from character-level/symbolic input can be effective until a certain point, where the relation can be inferred from explicit mentions in the text. Nevertheless, in Chapter 7 it has been shown that even strong learners of this type of patterns, i.e., current LM based transformer models, are vulnerable to slight changes in the textual input and can benefit from adding semantic or discourse information into the training process. Specifically for AM, the integration of discourse knowledge was shown to be beneficial [142, 172]. Further, the performance of these models was shown to be unreliably for cases where an understanding of concepts, procedures and their interactions in the world is presupposed by the speaker/writer, such as for causal reasoning or argument comprehension. Human communication is based on agreed facts about the world, e.g., that a water body of a certain size is not a lake anymore, but a sea or an ocean. These facts about the world can only be partially learned from text alone in an unsupervised way by current state-of-the-art NLP models. Moreover, these facts function as warrants in scientific argumentation and are mostly implicitly presumed in the biomedical domain [161]. Thus, I suggest that this knowledge has to be induced externally. For instance, for the application of AM in the medical domain, the argumentative relation classification module could be revamped in future work to integrate expert domain knowledge and explicitly model presumed warrants, which could infuse reasons into the learning process.

These could be established interrelations of medical concepts in form of ontologies or knowledge bases. For example, that *quality of life* has various subscales, one of which can be *incontinence*, which lowers the *quality of life*. Thus, it would help in the prediction of relations when an evidence talks about *incontinence* and the claim about *quality of life*. Surely Language Model proponents would argue that this can be all learned from enough textual data. However, this information must first be in the raw data, second it needs to be encoded properly into the model with all its implicit consequences (which is where current models struggle and this is why I think it has to be added from an external source), and third fetched in the right context of application. The right meaning for the right context in the latter part is supposed to be usually determined by enhancing certain signals during fine-tuning, which is a supervised task. Thus, if the information is not present in the training data for fine-tuning, and the model is not able to abstract it somehow otherwise, the signal might be too weak to impact the classification. Furthermore, an ontological or knowledge base approaches could be exploited with regards to the explainability of the implicit reasoning process involved in decision of the classifier. For example, Green [161] proposed to model scientific publications as knowledge bases and derive arguments from these KBs based on argumentation schemes formulated in the logical programming language Prolog, and Longo and Hederman [10] justify their work with knowledge bases for the prediction of breast cancer recurrence with the obtained explainability of the decision. While these, especially the former argumentation schemes, deliberately model the detailed relationship between arguments, it necessitates a reliable formalization of unstructured text into knowledge bases, similar to the need for structured data of the decision support frameworks presented in Chapter 9.1.1. However, the output of such an approach would be explainable, which is definitely desirable as a the long term goal.

Additionally, the point of an enriched model for relationships between arguments becomes even more decisive when expanding the relation prediction from the intra-argument level, as it was done in this thesis, to the inter-argument level, i.e., relations between multiple trials. While in theory the existing intra-argument relation models can be applied also for the inter-argument relation classification, the results would probably be worse. As shown in Chapter 4, the proposed models had troubles learning a good representation of the attack relations resulting in the highest rate of misclassification. Inter-argument relations are naturally more controversial. Thus, they contain more contradicting/attacking relations, which proper identification suffers from the insufficiently learned representation of such. Furthermore, as has been also proposed by [172], the combinations of the components should be limited to reduce the numbers of false positives, which is naturally increased when classifying all possible combination of all argument components across turns/documents. In general, the inter-argument relation prediction task can be thought of as a user guided clustering tool of arguments about the same disease with the aim to automatically identify, for instance, possible controversies among the conclusions of

multiple RCTs evaluating related treatments for a certain disease. To develop inter-argument relation annotations to evaluate the model performance, the underlying clinical trials must examine the same interventions for the same disease. To this end, the AbstRCT was probed to estimate the feasibility of inter-argument annotations. Given that the neoplasm topic was selected for its high diversity, the direct drawback is its scarcity of trials with similar setups. Even the relatively narrow glaucoma subset does not contain enough similar trials to have a sizeable amount of annotations to properly evaluate Machine Learning models. However, this leaves this topic as a far-reaching future research direction together with the extension of the analysis to the full text of clinical studies. The latter extension would provide the capability to better capture inconsistencies within one single trial. As it has been noticed in the literature [177], sometimes RCT abstracts contain a more positive reporting of the main findings of the article than what stated in the full text. Employing Argument Mining methods to automatically identify these instances of misrepresentation and distortion of the results in RCTs is a challenging and crucial research line for healthcare intelligent applications.

Furthermore, my work has shown that there are synergies between the different extracting modules, such as evidence types, PICO elements or effect on outcomes. Future work could go into the direction of further integrating other methods to distill even more information. For instance, more fine-grained types of claims could be beneficial, similar to how it was done for evidence in Chapter 5. A framework for claims in clinical trials, as it was for example proposed for other biomedical articles by Blake [163], would be beneficial for the qualitative evaluation of the arguments. Also with respect to the aforementioned inter-argument relations, the type of a claim can be of valuable information.

Overall, the work described in this thesis addresses only one of many facets in Evidence-based Medicine, i.e., the detection and preparation of evidence. The synthesis with other facets, similar to as it was done in this thesis with the integration of the PICO framework and outcome analysis, can broaden the area of application. In particular, future work can go further into the direction of interconnecting the provided evidence with existing tools or frameworks for the qualitative evaluation of it, such as RobotReviewer, the aforementioned claim framework or the argument-based decision support systems described in Section 9.1.2. Nevertheless, the process of searching, selecting, appraising and properly applying evidence in EBM is highly complex and thus, the (semi-)automated detection and evaluation of relevant evidence remains an important line of research in Evidence-based Medicine.

Bibliography

[1] D. L. Sackett and W. M. C. Rosenberg. "On the need for evidence-based medicine". In: *Journal of Public Health* 17.3 (1995), pp. 330–334.

[2] L. Manchikanti. "Evidence-Based Medicine, Systematic Reviews, and Guidelines in Interventional Pain Management Part I: Introduction and General Considerations". In: *Pain physician* 11.2 (2008), pp. 161–86.

[3] D. L. Sackett, W. M. Rosenberg, J. Gray, R. Haynes, and W. Richardson. "Evidence based medicine: What it is and what it isn't". In: *BMJ (Clinical research ed.)* 312 (1996), pp. 71–2.

[4] G. H. Guyatt. "Evidence-based medicine". In: *ACP Journal Club* (1991), A–16.

[5] A. Trenta, A. Hunter, and S. Riedel. "Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints". In: *CoRR* abs/1509.05209 (2015).

[6] B. Nye, J. J. Li, R. Patel, Y. Yang, I. Marshall, A. Nenkova, and B. Wallace. "A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*. Association for Computational Linguistics, 2018, pp. 197–207.

[7] D. Jin and P. Szolovits. "PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks". In: *Proceedings of the 17th Workshop on Biomedical Natural Language Processing (BioNLP 2018)*. 2018, pp. 67–75.

[8] A. Hunter and M. Williams. "Aggregating evidence about the positive and negative effects of treatments". In: *Artificial Intelligence in Medicine* 56.3 (2012), pp. 173–190.

[9] R. Craven, F. Toni, C. Cadar, A. Hadad, and M. Williams. "Efficient Argumentation for Medical Decision-Making". In: *Proceedings of 13th International Conference on the Principles of Knowledge Representation and Reasoning*. AAAI Press, 2012, pp. 598–602.

[10] L. Longo and L. Hederman. "Argumentation Theory for Decision Support in Health-Care: A Comparison with Machine Learning". In: *Proceedings of the International Conference on Brain and Health Informatics 2013*. Springer. 2013, pp. 168–180.

[11] M. Chary, S. Parikh, A. Manini, E. Boyer, and M. Radeous. "A Review of Natural Language Processing in Medical Education". In: *Western Journal of Emergency Medicine* 20.1 (2018), pp. 78–86.

[12] A. Peldszus and M. Stede. "From Argument Diagrams to Argumentation Mining in Texts: A Survey". In: *International Journal of Cognitive Informatics and Natural Intelligence* 7.1 (2013), pp. 1–31.

[13] M. Lippi and P. Torroni. "Argumentation Mining: State of the Art and Emerging Trends". In: *ACM Transactions on Internet Technology* 16.2 (2016), pp. 1–25.

[14] E. Cabrio and S. Villata. "Five Years of Argument Mining: a Data-driven Analysis". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*. International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 5427–5433.

[15] J. Lawrence and C. Reed. "Argument Mining: A Survey". In: *Computational Linguistics* 45.4 (2020), pp. 765–818.

[16] N. Green. "Argumentation for Scientific Claims in a Biomedical Research Article". In: *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing (ArgNLP 2014)*. Vol. 1341. CEUR Workshop Proceedings. 2014, pp. 5–10.

[17] S. Teufel. "Scientific Argumentation Detection as Limited-domain Intention Recognition". In: *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing (ArgNLP 2014)*. Vol. 1341. CEUR Workshop Proceedings. 2014, pp. 101–109.

[18] N. L. Green. "Annotating evidence-based argumentation in biomedical text". In: *Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2015)*. IEEE. 2015, pp. 922–929.

[19] M. A. Qassas, D. Fogli, M. Giacomin, and G. Guida. "Analysis of Clinical Discussions Based on Argumentation Schemes". In: *Procedia Computer Science* 64 (2015). Conference on ENTERprise Information Systems/International Conference on Project MANagement/Conference on Health and Social Care Information Systems and Technologies, pp. 282–289.

[20] H. Hassanzadeh, M. Kholghi, A. N. Nguyen, and K. Chu. "Clinical Document Classification Using Labeled and Unlabeled Data Across Hospitals". In: *American Medical Informatics Association Annual Symposium*. AMIA, 2018, pp. 545–554.

[21] W. Yoon, J. Lee, D. Kim, M. Jeong, and J. Kang. "Pre-trained Language Model for Biomedical Question Answering". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2019)*. Vol. 1168. Communications in Computer and Information Science. Springer, 2019, pp. 727–740.

[22] J. Liang, C.-H. Tsou, and A. Poddar. "A Novel System for Extractive Clinical Note Summarization using EHR Data". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP 2019)*. Association for Computational Linguistics, 2019, pp. 46–54.

[23] P. Besnard, A. Garcia, A. Hunter, S. Modgil, H. Prakken, G. Simari, and F. Toni. "Introduction to structured argumentation". In: *Argument & Computation* 5.1 (2014), pp. 1–4.

[24] G. Guyatt et al. "Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine". In: *Journal of the American Medical Association* 268.17 (1992), pp. 2420–2425.

[25] W. Richardson, M. Wilson, J. Nishikawa, and R. Hayward. "The well-built clinical question: a key to evidence-based decisions." In: *ACP journal club* 123.3 (1995), A12–3.

[26] T. J. Bench-Capon and P. E. Dunne. "Argumentation in artificial intelligence". In: *Artificial Intelligence* 171.10 (2007), pp. 619–641.

[27] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. R. Simari, M. Thimm, and S. Villata. "Towards Artificial Argumentation". In: *AI Magazine* 38.3 (2017), pp. 25–36.

[28] S. Teufel, A. Siddharthan, and C. Batchelor. "Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*. 2009, pp. 1493–1502.

[29] M.-F. Moens, E. Boiy, R. M. Palau, and C. Reed. "Automatic Detection of Arguments in Legal Texts". In: *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL 2007)*. Association for Computing Machinery, 2007, pp. 225–230.

[30] R. Mochales and M.-F. Moens. "Argumentation mining". In: *Artificial Intelligence and Law* 19.1 (2011), pp. 1–22.

[31] E. Cabrio and S. Villata. "A Natural Language Bipolar Argumentation Approach to Support Users in Online Debate Interactions". In: *Argument & Computation* 4.3 (2013), pp. 209–230.

[32] F. Boltužić and J. Šnajder. "Back up your Stance: Recognizing Arguments in Online Discussions". In: *Proceedings of the 1st Workshop on Argumentation Mining (ArgMining 2014)*. Association for Computational Linguistics, 2014, pp. 49–58.

[33] I. Habernal, J. Eckle-Kohler, and I. 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 (ArgNLP 2014)*. Vol. 1341. CEUR Workshop Proceedings. 2014, pp. 26–39.

[34] D. N. Walton. *Argumentation schemes for presumptive reasoning*. Psychology Press, 1996.

[35] D. Küçük and F. Can. "Stance Detection: A Survey". In: *ACM Computing Surveys* 53.1 (2020), pp. 1–37.

[36] S. Eger, J. Daxenberger, and I. Gurevych. "Neural End-to-End Learning for Computational Argumentation Mining". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*. Association for Computational Linguistics, 2017, pp. 11–22.

[37] G. Morio and K. Fujita. "End-to-End Argument Mining for Discussion Threads Based on Parallel Constrained Pointer Architecture". In: *Proceedings of the 5th Workshop on Argument Mining (ArgMining 2018)*. Association for Computational Linguistics, 2018, pp. 11–21.

[38] N. Reimers, B. Schiller, T. Beck, J. Daxenberger, C. Stab, and I. Gurevych. "Classification and Clustering of Arguments with Contextualized Word Embeddings". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics, 2019, pp. 567–578.

[39] H. Wachsmuth, B. Stein, and Y. Ajjour. "PageRank for argument relevance". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. Association for Computational Linguistics, 2017, pp. 1117–1127.

[40] A. Toledo, S. Gretz, E. Cohen-Karlik, R. Friedman, E. Venezian, D. Lahav, M. Jacovi, R. Aharonov, and N. Slonim. "Automatic Argument Quality Assessment - New Datasets and Methods". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, 2019, pp. 5625–5635.

[41] L. Gienapp, B. Stein, M. Hagen, and M. Potthast. "Efficient Pairwise Annotation of Argument Quality". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Association for Computational Linguistics, 2020, pp. 5772–5781.

[42] J. Lawrence, J. Visser, and C. Reed. "Harnessing rhetorical figures for argument mining". In: *Argument & Computation* 8.3 (2017), pp. 289–310.

[43] T. Mayer, E. Cabrio, and S. Villata. "Evidence Type Classification in Randomized Controlled Trials". In: *Proceedings of the 5th Workshop on Argument Mining (ArgMining 2018)*. Association for Computational Linguistics, 2018, pp. 29–34.

[44] K. S. Jones. "Natural language processing: a historical review". In: *University of Cambridge* (2001), pp. 2–10.

[45] N. Indurkhya and F. J. Damerau. *Handbook of natural language processing*. Vol. 2. CRC Press, 2010.

[46] W. J. Hutchins. "Machine translation: A brief history". In: *Concise history of the language sciences*. Elsevier, 1995, pp. 431–445.

[47] N. Chomsky. *On nature and language*. Cambridge University Press, 2002.

[48] S. T. Piantadosi, H. Tily, and E. Gibson. "The communicative function of ambiguity in language". In: *Cognition* 122.3 (2012), pp. 280–291.

[49] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space". In: *Workshop Track Proceedings of the 1st International Conference on Learning Representations (ICLR 2013)*. 2013.

[50] J. Pennington, R. Socher, and C. D. Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Association for Computational Linguistics, 2014, pp. 1532–1543.

[51] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. "Learning Word Vectors for 157 Languages". In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association, 2018, pp. 3483–3487.

[52] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*. Association for Computational Linguistics, 2018, pp. 2227–2237.

[53] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[54] A. Akbik, D. Blythe, and R. Vollgraf. "Contextual String Embeddings for Sequence Labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2028)*. Association for Computational Linguistics, 2018, pp. 1638–1649.

[55] J. Howard and S. Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*. Association for Computational Linguistics, 2018, pp. 328–339.

[56] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies". In: *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*. Curran Associates, Inc., 2017, pp. 6000–6010.

[58] Y. Wu et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: *CoRR* abs/1609.08144 (2016).

[59] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[60] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP 2018)*. Association for Computational Linguistics, 2018, pp. 353–355.

[61] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. "Improving language understanding by generative pre-training". In: *OpenAI report* (2018).

[62] W. L. Taylor. ""Cloze Procedure": A New Tool for Measuring Readability". In: *Journalism Quarterly* 30.4 (1953), pp. 415–433.

[63] I. Beltagy, K. Lo, and A. Cohan. "SciBERT: A Pretrained Language Model for Scientific Text". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, 2019, pp. 3615–3620.

[64] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[65] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Association for Computational Linguistics, 2020, pp. 7203–7219.

[66] G. Lample and A. Conneau. "Cross-lingual Language Model Pretraining". In: *Proceedings of the 33st International Conference on Neural Information Processing Systems (NeurIPS 2019)*. Curran Associates, Inc., 2019, pp. 7059–7069.

[67] Y. Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019).

[68] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. "SpanBERT: Improving Pre-training by Representing and Predicting Spans". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 64–77.

[69] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang. "ERNIE 2.0: A Continual Pre-training Framework for Language Understanding". In: *CoRR* abs/1907.12412 (2019).

[70] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou. "Semantics-aware BERT for language understanding". In: *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*. AAAI Press, 2020, pp. 9628–9635.

[71] T. Mihaylov and A. Frank. "Discourse-Aware Semantic Self-Attention for Narrative Reading Comprehension". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, 2019, pp. 2541–2552.

[72] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. "ERNIE: Enhanced Language Representation with Informative Entities". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics, 2019, pp. 1441–1451.

[73] I. Tenney, D. Das, and E. Pavlick. "BERT Rediscovers the Classical NLP Pipeline". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics, 2019, pp. 4593–4601.

[74] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. "What Does BERT Look at? An Analysis of BERT's Attention". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP 2019)*. Association for Computational Linguistics, 2019, pp. 276–286.

[75] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *8th International Conference on Learning Representations (ICLR 2020)*. OpenReview.net, 2020.

[76] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *CoRR* abs/1910.01108 (2019).

[77] N. Kitaev, L. Kaiser, and A. Levskaya. "Reformer: The Efficient Transformer". In: *8th International Conference on Learning Representations (ICLR 2020)*. OpenReview.net, 2020.

[78] E. Strubell, A. Ganesh, and A. McCallum. "Energy and Policy Considerations for Deep Learning in NLP". In: *CoRR* abs/1906.02243 (2019).

[79] T. Mayer, E. Cabrio, M. Lippi, P. Torroni, and S. Villata. "Argument Mining on Clinical Trials". In: *Proceedings of 7th International Conference on Computational Models of Argument (COMMA 2018)*. IOS Press, 2018, pp. 137–148.

[80] T. Mayer, E. Cabrio, and S. Villata. "Transformer-Based Argument Mining for Healthcare Applications". In: *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*. IOS Press, 2020, pp. 2108–2115.

[81] F. Dernoncourt, J. Y. Lee, and P. Szolovits. "Neural Networks for Joint Sentence Classification in Medical Paper Abstracts". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. Association for Computational Linguistics, 2017, pp. 694–700.

[82] E. Hannan. "Randomized Clinical Trials and Observational Studies Guidelines for Assessing Respective Strengths and Limitations". In: *JACC. Cardiovascular interventions* 1.3 (2008), pp. 211–217.

[83] G. H. Guyatt et al. "Users' Guides to the Medical Literature XXV. Evidence-Based Medicine: Principles for Applying the Users' Guides to Patient Care". In: *Journal of the American Medical Association* 284.10 (2000), pp. 1290–1296.

[84] K. F. Schulz and D. A. Grimes. "Generation of allocation sequences in randomised trials: chance, not choice". In: *The Lancet* 359.9305 (2002), pp. 515–519.

[85] C. Stab and I. Gurevych. "Parsing Argumentation Structures in Persuasive Essays". In: *Computational Linguistics* 43.3 (2017), pp. 619–659.

[86] S. E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2003.

[87] L. A. Groarke and C. W. Tindale. *Good reasoning matters: A constructive approach to critical thinking*. Oxford University Press, 1997.

[88] G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. "The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts". In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP 2008)*. Association for Computational Linguistics, 2008, pp. 38–45.

[89] E. Lehman, J. DeYoung, R. Barzilay, and B. C. Wallace. "Inferring Which Medical Treatments Work from Reports of Clinical Trials". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. Association for Computational Linguistics, 2019, pp. 3705–3717.

[90] S. Hopewell, M. Clarke, D. Moher, E. Wager, P. Middleton, D. G. Altman, K. F. Schulz, and the CONSORT Group. "CONSORT for Reporting Randomized Controlled Trials in Journal and Conference Abstracts: Explanation and Elaboration". In: *PLOS Medicine* 5.1 (2008), pp. 1–9.

[91] J. L. Fleiss. "Measuring nominal scale agreement among many raters". In: *Psychological bulletin* 76.5 (1971), pp. 378–382.

[92] A. Zapf, S. Castell, L. Morawietz, and A. Karch. "Measuring inter-rater reliability for nominal data - Which coefficients and confidence intervals are appropriate?" In: *BMC Medical Research Methodology* 16.1 (2016), pp. 93–103.

[93] K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[94] J. R. Landis and G. G. Koch. "The measurement of observer agreement for categorical data". In: *Biometrics* 33.1 (1977), pp. 159–174.

[95] M. Lippi and P. Torroni. "MARGOT: A web server for argumentation mining". In: *Expert Systems with Applications* 65 (2016), pp. 292–303.

[96] E. Aharoni, A. Polnarov, T. Lavee, D. Hershcovich, R. Levy, R. Rinott, D. Gutfreund, and N. Slonim. "A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics". In: *Proceedings of the 1st Workshop on Argumentation Mining (ArgMining 2014)*. Association for Computational Linguistics, 2014, pp. 64–68.

[97] R. Rinott, L. Dankin, C. A. Perez, M. M. Khapra, E. Aharoni, and N. Slonim. "Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Association for Computational Linguistics, 2015, pp. 440–450.

[98] M. Collins and N. Duffy. "New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*. ACL, 2002, pp. 263–270.

[99] M. Lippi and P. Torroni. "Context-Independent Claim Detection for Argument Mining". In: *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*. AAAI Press, 2015, pp. 185–191.

[100] A. Moschitti. "State-of-the-art Kernels for Natural Language Processing". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, 2012, pp. 2–2.

[101] A. Moschitti. "Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees". In: *Proceedings of the 17th European Conference on Machine Learning (ECML 2006)*. Springer, 2006, pp. 318–329.

[102] A. Komninos and S. Manandhar. "Dependency Based Embeddings for Sentence Classification Tasks". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*. Association for Computational Linguistics, 2016, pp. 1490–1500.

[103] B. Heinzerling and M. Strube. "BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages". In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association, 2018, pp. 2989–2993.

[104] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *CoRR* abs/1406.1078 (2014).

[105] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*. Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[106] M. Miwa and M. Bansal. "End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Association for Computational Linguistics, 2016, pp. 1105–1116.

[107] A. Galassi, M. Lippi, and P. Torroni. "Argumentative Link Prediction using Residual Networks and Multi-Objective Learning". In: *Proceedings of the 5th Workshop on Argument Mining (ArgMining 2018)*. Association for Computational Linguistics, 2018, pp. 29–34.

[108] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi. "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Association for Computational Linguistics, 2018, pp. 93–104.

[109] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard. "Complex Embeddings for Simple Link Prediction". In: *Proceedings of The 33rd International Conference on Machine Learning (ICML 2016)*. PMLR, 2016, pp. 2071–2080.

[110] I. Balazevic, C. Allen, and T. Hospedales. "TuckER: Tensor Factorization for Knowledge Graph Completion". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, 2019, pp. 5185–5194.

[111] A. Bordes, N. Usunier, A. Garcıa-Durán, J. Weston, and O. Yakhnenko. "Translating Embeddings for Modeling Multi-relational Data". In: *Proceedings of 27th Conference on Neural Information Processing Systems (NeurIPS 2013)*. Curran Associates Inc., 2013, pp. 2787–2795.

[112] T. Dettmers, M. Pasquale, S. Pontus, and S. Riedel. "Convolutional 2D Knowledge Graph Embeddings". In: *Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI 2018)*. AAAI Press, 2018, pp. 1811–1818.

[113] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein. "SemEval-2018 Task 12: The Argument Reasoning Comprehension Task". In: *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval 2018)*. Association for Computational Linguistics, 2018, pp. 763–772.

[114] R. Bellomo and S. M. Bagshaw. "Evidence-based medicine: Classifying the evidence from clinical trials – the need to consider other dimensions". In: *Critical Care* 10.5 (2006), pp. 232–240.

[115] D. H. Park and C. Blake. "Identifying Comparative Claim Sentences in Full-text Scientific Articles". In: *Proceedings of the 2012 ACL Workshop on Detecting Structure in Scholarly Discourse (DSSD 2012)*. Association for Computational Linguistics, 2012, pp. 1–9.

[116] S. Gupta, A. S. M. A. Mahmood, K. Ross, C. H. Wu, and K. Vijay-Shanker. "Identifying Comparative Structures in Biomedical Text". In: *Proceeding of the 16th Workshop on Biomedical Natural Language Processing (BioNLP 2017)*. Association for Computational Linguistics, 2017, pp. 206–215.

[117] T. Mayer, S. Marro, E. Cabrio, and S. Villata. "Generating Adversarial Examples for Topic-Dependent Argument Classification". In: *Proceedings of 8th International Conference on Computational Models of Argument (COMMA 2020)*. IOS Press, 2020, pp. 33–44.

[118] T. Mayer. "Enriching Language Models with Semantics". In: *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*. IOS Press, 2020, pp. 2917–2918.

[119] C. Stab, T. Miller, B. Schiller, P. Rai, and I. Gurevych. "Cross-topic Argument Mining from Heterogeneous Sources". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Association for Computational Linguistics, 2018, pp. 3664–3674.

[120] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Association for Computational Linguistics, 2015.

[121] E. Shnarch, C. Alzate, L. Dankin, M. Gleize, Y. Hou, L. Choshen, R. Aharonov, and N. Slonim. "Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*. Association for Computational Linguistics, 2018, pp. 599–605.

[122] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. "Intriguing properties of neural networks". In: *2nd International Conference on Learning Representations (ICLR 2014)*. 2014.

[123] I. J. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and Harnessing Adversarial Examples". In: *3rd International Conference on Learning Representations (ICLR 2015)*. 2015.

[124] X. Yuan, P. He, Q. Zhu, R. R. Bhat, and X. Li. "Adversarial Examples: Attacks and Defenses for Deep Learning". In: *IEEE Transactions on Neural Networks and Learning Systems* 30.9 (2019), pp. 2805–2824.

[125] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li. "Adversarial Attacks on Deep-learning Models in Natural Language Processing: A Survey". In: *ACM Transactions on Intelligent Systems and Technology* 11.3 (2020), pp. 1–41.

[126] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer. "Adversarial Example Generation with Syntactically Controlled Paraphrase Networks". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*. Association for Computational Linguistics, 2018, pp. 1875–1885.

[127] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang. "Generating Natural Language Adversarial Examples". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Association for Computational Linguistics, 2018, pp. 2890–2896.

[128] Y.-L. Hsieh, M. Cheng, D.-C. Juan, W. Wei, W.-L. Hsu, and C.-J. Hsieh. "On the Robustness of Self-Attentive Models". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics, 2019, pp. 1520–1529.

[129] T. Niven and H.-Y. Kao. "Probing Neural Network Comprehension of Natural Language Arguments". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics, 2019, pp. 4658–4664.

[130] R. Jia and P. Liang. "Adversarial Examples for Evaluating Reading Comprehension Systems". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. Association for Computational Linguistics, 2017, pp. 2021–2031.

[131] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. "Practical Black-Box Attacks against Machine Learning". In: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ACM AsiaCCS 2017)*. Association for Computing Machinery, 2017, pp. 506–519.

[132] E. F. Tjong Kim Sang and F. De Meulder. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the 7th Conference on Natural Language Learning (CoNLL 2003)*. 2003, pp. 142–147.

[133] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67.

[134] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. "HellaSwag: Can a Machine Really Finish Your Sentence?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics, 2019, pp. 4791–4800.

[135] E. M. Bender and A. Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Association for Computational Linguistics, 2020, pp. 5185–5198.

[136] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Association for Computational Linguistics, 2020, pp. 4902–4912.

[137] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019)*. Curran Associates, Inc., 2019, pp. 3266–3280.

[138] M. Roemmele, C. Bejan, and A. Gordon. "Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning". In: *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*. 2011, pp. 90–95.

[139] H. J. Levesque, E. Davis, and L. Morgenstern. "The Winograd Schema Challenge". In: *Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning (KR 2012)*. AAAI Press, 2012, pp. 552–561.

[140] R. Speer and C. Havasi. "Representing General Relational Knowledge in ConceptNet 5". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association, 2012, pp. 3679–3686.

[141] M. Sap, R. LeBras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. "ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning". In: *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*. AAAI Press, 2019, pp. 3027–3035.

[142] F. Hewett, R. Prakash Rane, N. Harlacher, and M. Stede. "The Utility of Discourse Parsing Features for Predicting Argumentation Structure". In: *Proceedings of the 6th Workshop on Argument Mining (ArgMining 2019)*. Association for Computational Linguistics, 2019, pp. 98–103.

[143] T. Mayer, E. Cabrio, and S. Villata. "ACTA A Tool for Argumentative Clinical Trial Analysis". In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*. International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 6551–6553.

[144] F. Michel et al. "Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research". In: *Proceedings of the 19th International Semantic Web Conference (ISWC 2020)*. in-press, 2020.

[145] L. L. Wang et al. "CORD-19: The Covid-19 Open Research Dataset". In: *ArXiv* abs/2004.10706 (2020).

[146] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. "Improving efficiency and accuracy in multilingual entity extraction". In: *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTiCS 2013)*. Association for Computing Machinery, 2013, pp. 121–124.

[147] C. Jonquet, N. H. Shah, and M. A. Musen. "The open biomedical annotator". In: *Summit on Translational Bioinformatics* 2009 (2009), p. 56.

[148] M. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (2016), pp. 1–9.

[149] O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker. "Querying the Semantic Web with Corese Search Engine". In: *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*. IOS Press, 2004, pp. 705–709.

[150] R. A. Cava, C. M. D. S. Freitas, and M. Winckler. "ClusterVis: visualizing nodes attributes in multivariate graphs". In: *Proceedings of the 32nd Symposium on Applied Computing (SAC 2017)*. ACM, 2017, pp. 174–179.

[151] M. Bersanelli. "Controversies about COVID-19 and Anticancer Treatment with Immune Checkpoint Inhibitors". In: *Immunotherapy* 12.5 (2020), pp. 269–273.

[152] M. Neumann, D. King, I. Beltagy, and W. Ammar. "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing". In: *Proceedings of the 18th Workshop on Biomedical Natural Language Processing (BioNLP 2019)*. Association for Computational Linguistics, 2019, pp. 319–327.

[153] I. Rahwan, F. Zablith, and C. Reed. "Laying the foundations for a World Wide Argument Web". In: *Artificial Intelligence* 171.10 (2007), pp. 897–921.

[154] M. Williams, Z. Liu, A. Hunter, and F. Macbeth. "An updated systematic review of lung chemo-radiotherapy using a new evidence aggregation method". In: *Lung Cancer* 87 (2014), pp. 290–295.

[155] P. M. Dung, R. A. Kowalski, and F. Toni. "Assumption-Based Argumentation". In: *Argumentation in Artificial Intelligence*. Springer, 2009, pp. 199–218.

[156] M. Fiszman, D. Demner-Fushman, F. Lang, P. Goetz, and T. C. Rindflesch. "Interpreting comparative constructions in biomedical text". In: *Proceeding of the 2007 ACL Workshop on Biological, Translational, and Clinical Language Processing (BioNLP 2007)*. Association for Computational Linguistics, 2007, pp. 137–144.

[157] O. Zaidan, J. Eisner, and C. Piatko. "Using "Annotator Rationales" to Improve Machine Learning for Text Categorization". In: *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*. Association for Computational Linguistics, 2007, pp. 260–267.

[158] S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, and I. Sim. "ExaCT: Automatic extraction of clinical trial characteristics from journal publications". In: *BMC medical informatics and decision making* 10 (2010), p. 56.

[159] I. J. Marshall, J. Kuiper, and B. C. Wallace. "RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials". In: *Journal of the American Medical Informatics Association* 23.1 (2016), pp. 193–201.

[160] I. Marshall, J. Kuiper, E. Banner, and B. C. Wallace. "Automating Biomedical Evidence Synthesis: RobotReviewer". In: *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, 2017, pp. 7–12.

[161] N. L. Green. "Towards mining scientific discourse using argumentation schemes". In: *Argument & Computation* 9.2 (2018), pp. 121–135.

[162] A. Alamri and R. Stevenson. "A Corpus of Potentially Contradictory Research Claims from Cardiovascular Research Abstracts". In: *Journal of Biomedical Semantics* 7.1 (2016), pp. 36–45.

[163] C. Blake. "Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles". In: *Biomedical Informatics* 43.2 (2010), pp. 173–189.

[164] R. Bar-Haim, I. Bhattacharya, F. Dinuzzo, A. Saha, and N. Slonim. "Stance Classification of Context-Dependent Claims". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. Association for Computational Linguistics, 2017, pp. 251–261.

[165] S. Menini, E. Cabrio, S. Tonelli, and S. Villata. "Never Retreat, Never Retract: Argumentation Analysis for Political Speeches". In: *Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI 2018)*. AAAI Press, 2018, pp. 4889–4896.

[166] X. Hua, M. Nikolov, N. Badugu, and L. Wang. "Argument Mining for Understanding Peer Reviews". In: *"Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. Association for Computational Linguistics, 2019, pp. 2131–2137.

[167] A. Søgaard and Y. Goldberg. "Deep multi-task learning with low level tasks supervised at lower layers". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Association for Computational Linguistics, 2016, pp. 231–235.

[168] Y. Ajjour, W.-F. Chen, J. Kiesel, H. Wachsmuth, and B. Stein. "Unit Segmentation of Argumentative Texts". In: *Proceedings of the 4th Workshop on Argument Mining (ArgMining 2017)*. Association for Computational Linguistics, 2017, pp. 118–128.

[169] M. Spliethöver, J. Klaff, and H. Heuer. "Is It Worth the Attention? A Comparative Evaluation of Attention Layers for Argument Unit Segmentation". In: *Proceedings of the 6th Workshop on Argument Mining (ArgMining 2019)*. Association for Computational Linguistics, 2019, pp. 74–82.

[170] P. Potash, A. Romanov, and A. Rumshisky. "Here's My Point: Joint Pointer Architecture for Argument Mining". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. Association for Computational Linguistics, 2017, pp. 1364–1373.

[171] V. Niculae, J. Park, and C. Cardie. "Argument Mining with Structured SVMs and RNNs". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*. Association for Computational Linguistics, 2017, pp. 985–995.

[172] T. Chakrabarty, C. Hidey, S. Muresan, K. McKeown, and A. Hwang. "AMPERSAND: Argument Mining for PERSuAsive oNline Discussions". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, 2019, pp. 2933–2943.

[173] M. Liakata, P. Thompson, A. Waard, R. Nawaz, H. Maat, and S. Ananiadou. "A three-way perspective on scientific discourse annotation for knowledge extraction". In: *Proceedings of the 2012 ACL Workshop on Detecting Structure in Scholarly Discourse (DSSD 2012)*. Association for Computational Linguistics, 2012, pp. 37–46.

[174] M. Liakata, S. Saha, S. Dobnik, C. R. Batchelor, and D. Rebholz-Schuhmann. "Automatic recognition of conceptualization zones in scientific articles and two life science applications". In: *Bioinformatics* 28.7 (2012), pp. 991–1000.

[175] C. Kirschner, J. Eckle-Kohler, and I. Gurevych. "Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications". In: *Proceedings of the 2nd Workshop on Argument Mining (ArgMining 2015)*. The Association for Computational Linguistics, 2015, pp. 1–11.

[176] W. Mann and S. Thompson. "Rethorical Structure Theory: Toward a functional theory of text organization". In: *Text-Interdisciplinary Journal for the Study of Discourse* 8.3 (1988), pp. 243–281.

[177] I. Boutron and P. Ravaud. "Misrepresentation and distortion of research in biomedical literature". In: *Proceedings of the National Academy of Sciences* 115.11 (2018), pp. 2613–2619.