A Survey on how to Cross-Reference Web Information Sources

Joe Raad, Aurélie Bertaux, Christophe Cruz

A Survey on how to Cross-Reference Web Information Sources

Joe Raad[1], Aurelie Bertaux[2], and Christophe Cruz[3]

CheckSem Department, Le2i Laboratory
University of Burgundy
Dijon, France
[1]Joe_Raad@etu.u-bourgogne.fr, [2]Aurelie.Bertaux@u-bourgogne.fr, [3]Christophe.Cruz@u-bourgogne.fr

***Abstract*—The goal of giving information a well-defined meaning is currently shared by different research communities. Once information has a well-defined meaning, it can be searched and retrieved more effectively. Therefore, this paper is a survey about the methods that compare different textual information sources in order to determine whether they address a similar information or not. The improvement of the studied methods will eventually lead to increase the efficiency of documentary research. In order to achieve this goal, the first category of methods focuses on semantic measure definitions. A second category of methods focuses on paraphrase identification techniques, and the last category deals with improving event extraction techniques. A general discussion is given at the end of the paper presenting the advantages and disadvantages of these methods.**

Keywords—Cross-Reference Web Information Sources; Documentary Research; Event Extraction; Paraphrase Identification; Semantic Measures; Semantic Relatedness; Similarity Definition.

I. INTRODUCTION

The quantity of the data on the web is increasing day after day in an immense way. This property is mainly due to the ease of creating web sites and the importance of social medias. In fact, and according to the analysis of Lawrence and Giles [31], the page number of World Wide Web doubles every two years. As a result, the difficulty in analyzing and retrieving information on the web increases with the growth of the number of pages. This problem proves the necessity of integrating knowledge in the web. This leads us to the purpose of converting the current web, dominated by unstructured and semi-structured documents into a "web of data" that can be processed and analyzed by machines.

This paper focuses on one aspect, which is how to improve cross-referencing of web information sources. This basically means linking the different textual information sources that share similar meanings. In order to reach this goal, one must focus on extracting knowledge, and precisely *extracting events* from different web information sources. However, extracting useful structured representation of events from a disorganized source like the web is a challenging problem. As one can express a single event in thousands of ways in natural language sentences. Therefore, *paraphrase identification*, which determines whether or not two formally distinct strings are similar in meaning is an effective way to solve this problem. Furthermore, in order to identify whether two expressions are paraphrases or not, these techniques rely on evaluating the *semantic measure* between these expressions.

Next section presents the definition and general concepts of semantic measure. Section three deals with the different approaches of cross-referencing information by discussing the following main approaches: *Semantic measure* definitions, *paraphrase identification* and *event extraction*. Finally, before concluding, the last section discusses the limits and the complementarity of these methods.

II. CONTEXT

In order to compare cross-referencing methods of web information sources, one must try to evaluate the Semantic Measure between concepts. Semantic Measure is normally a philosophic term. It is a point of view which differs from one person to another regarding the semantic links strengths between two concepts. Trying to computerize this philosophic term, in order to compare different textual information, is a complex task and requires to perform high-level language processing. However, the evaluation of the Semantic Measure between two concepts depends firstly on the type of the semantic links, and secondly on the type of knowledge resources.

A. Semantic Measure Types

In order to compare two concepts, and precisely two textual information sources in the case of documentary research, one must evaluate the semantic measure between these sources. However, semantic measure is a generic term covering several concepts:

- **Semantic relatedness**, which is the most general semantic link between two concepts. Two concepts do not have to share a common meaning to be considered semantically related or close, they can be linked by a functional relationship or frequent

association relationship like meronym or antonym concepts.
(e.g. Pilot "is related to" Airplane)

- **Semantic similarity**, which is a specific case of semantic relatedness. Two concepts are considered similar if they share common meanings and characteristics, like synonym, hyponym and hypernym concepts. (e.g. Old "is similar to" Ancient)
- **Semantic distance**, is the inverse of the semantic relatedness, as it indicates how much two concepts are unrelated to one another.

B. Knowledge Resources Types

The evaluation of the semantic measure strongly depends on the level of knowledge of the person comparing different concepts, which is the knowledge resource in the case of a machine trying to perform this task. For instance, the relationship of the two terms "Titanic" and "Avatar" does not exist at all for a given person. But, another person identifies them as related since these terms are both movie titles. Furthermore, a movie addict strongly relates these two terms, as they are not only movie titles, but these movies also share the same writer and director. Therefore, we can see the influence and the importance of the knowledge resources (level of knowledge for humans) in the evaluation of the semantic measure between two concepts.

Different knowledge sources are used in different approaches. Some of the approaches use only one knowledge source, while other approaches use several ones. Among the most common sources, we can find structured sources like *ontologies* and the *DBPedia*, semi-structured sources like *Wikipedia* and unstructured sources like textual data from the *web*.

1) Structured Sources

Ontologies

The word *ontology* is generally used to mean different things, e.g. glossaries & data dictionaries, thesauri & taxonomies, schemas & data models, and formal ontologies & inference. These different databases have the same goal, to provide information on the meaning of elements. One of the most used ontologies is the large English lexical database WordNet. In WordNet, there are four commonly used semantic relations for nouns, which are hyponym/hypernym (is-a), part meronym/part holonym (part-of), member meronym/member holonym (member-of) and substance meronym/substance holonym (substance-of). A fragment of (is-a) relation between concepts in WordNet is shown in Figure1. We can also find many other popular general purpose ontologies like YAGO and SENSUS, and some domain specific ontologies like UMLS and MeSH (for biomedical and health related concepts), SNOMED (for clinical healthcare concepts), GO (for gene proteins and all concerns of organisms) and STDS (for earth-referenced spatial data).

However, the use of ontologies shows some limitations in their low coverage (e.g. containing few proper names) and in their need for experts to supply their knowledge, which makes it difficult to be expanded and updated.

DBPedia

DBPedia is a project aiming to extract structured contents from Wikipedia, by allowing users to query sophisticated queries containing relationships and properties. Its structure combined with a very large amount of data, makes DBPedia a reliable and important source of knowledge for several applications including semantic measurements.

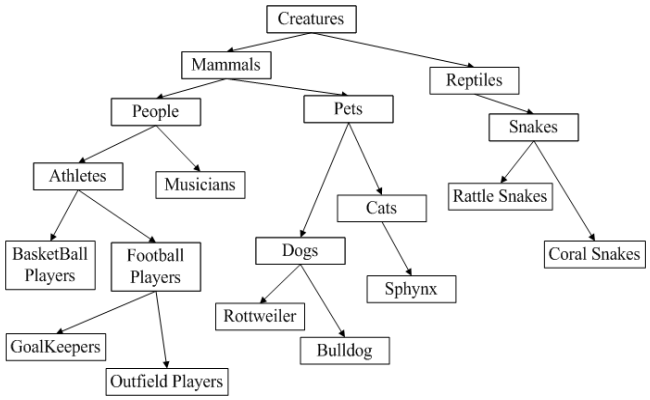

Figure 1. A Fragment of (is-a) Relation in WordNet

2) Semi-Structured Sources

Wikipedia

Wikipedia is the most popular internet encyclopedia, and the seventh most popular website in the world according to Alexa's latest web data analysis [32]. Therefore, being the Internet's largest and most popular general reference work makes Wikipedia one of the most used knowledge resource in this field. Though the fact that it covers most of the fields and domains, and contains lot of proper nouns. It is not as important as the web to discover and evaluate semantic relatedness. Actually you cannot always find semantic related terms in the same Wikipedia article.

3) Unstructured Sources

Web

The web is now the biggest and the fastest growing source of information on earth. Therefore, it is one of the most important knowledge resource. Even though the number of experts is small compared to the total number of internet users, which can make it sometimes an unreliable source of information, it can still be easy to remove the noise in order to extract relevant information.

4) Discussion

From Table I, which represents a summary of the characteristics of the knowledge resources mentioned in this section, we can derive that every resource has its own different characteristics. Therefore, we cannot conclude which one is the best resource in general, but we can conclude which one is the most recommended in a specific use. For instance, if the high amount of information, the wide coverage and the high growth rate are our interest, it is recommended to use the web as knowledge resource. On the other hand, if the use of a structured source is our interest, it is recommended to use ontologies or DBPedia as our knowledge resource. And finally, if an average of all the characteristics is demanded, Wikipedia or also DBPedia are the best fitting solutions in this case.

In the next section, we briefly discuss some of the proposed methods working on improving *semantic measures*, *paraphrase identification* and *event extraction*.

TABLE I. *An Overview of Knowledge Resources*

Knowledge Source	***Size***	***Coverage***	***Structure***	***Growth***
Ontologies	*Low*	*Low*	*High*	*Low*
Wikipedia	*Medium*	*Medium*	*Medium*	*Medium*
DBPedia	*Medium*	*Medium*	*High*	*Medium*
Web	*High*	*High*	*Low*	*High*

In the next section, we briefly discuss some of the proposed methods working on improving *semantic measures*, *paraphrase identification* and *event extraction*.

III. Cross-Referencing Methods Proposals

Three sections presents the main cross-referencing methods for web information sources:

A. Semantic Measures

In order to improve cross-referencing methods for web information sources, semantic measures, and precisely semantic similarity definitions are proposed. These measures can normally be grouped into five categories: *Path Length-based measures, Information Content-based measures, Feature-based measures, Distributional-based measures and Hybrid measures*.

1) Path Length based measures

All approaches based on this type of measure, consider that the similarity between two concepts depends on the length of the path linking these two concepts and the positions in the taxonomy. In path length-based measures, the concepts are represented by nodes and the relationship between the concepts are represented by edges.

One basic approach is the *shortest path-based* measure, which consists in calculating the similarity measure based on the shortest path length between two concepts c_i, as in Eq. 1. For instance, and referring back to Figure 1, the similarity measure between the concepts "People" and "Pets" would be equal to the similarity measure between "Rottweiler" and "Bulldog", because they share the same length between each other.

$$\mathrm{sim}(c1, c2) = 2 * \mathrm{Max}_{\mathrm{depth}} - \mathrm{Length}(c1, c2) \qquad (1)$$

**Max_depth: The maximum depth(c_i) of the taxonomy.*
**Length(c_i,c_j): The length of the shortest path from c_i to c_j.*

Another similarity definition introduced by Leacock and Chodorow [26], also depends on the shortest path length between two concepts as shown in Eq. 2. Therefore, the same results are obtained as the previous example.

$$\mathrm{sim}(c1, c2) = -\log \frac{\mathrm{Length}(c1, c2)}{2 * \mathrm{Max\_depth}} \qquad (2)$$

In addition, Wu and Palmer [27] introduces a new similarity definition, that takes into account the position of the most specific common parent of the two concepts. Therefore the similarity between two concepts depends not only on the path length between these concepts, but also on the depth of their lowest common parent in the taxonomy as shown in Eq. 3. In this case, sim(People, Pets) is different than sim(Rottweiler, Bulldog) as the depths of their respective common parent "Mammals" and "Dogs" are different.

$$\mathrm{sim}(c1, c2) = \frac{2 * \mathrm{Depth}(\mathrm{Iso}(c1,c2))}{\mathrm{Length}(c1,c2) + 2 * \mathrm{Depth}(\mathrm{Iso}(c1,c2))} \qquad (3)$$

**Depth(c_i): The length from c_i to the global root concept. (Depth(root)=1)*
**Iso(c_i,c_j): The lowest common parent of c_i and c_j.*

2) Information Content based measures

In this kind of similarity measure, all approaches measure similarity based on the information content of each concept. The more common information two concepts share, the more similar these concepts are.

This method of measuring similarity is used by Resnik [8], who introduced in 1995 a similarity definition based on the notion of information content, using just taxonomic "is-a" links in general, with the addition of some other links such as "part-of". This approach assumes that for two given concepts, similarity depends on the information content that subsumes them in the taxonomy. Therefore in a hierarchy, the more general (the closest to root) the first common parent of two concepts is, the lower the similarity between them. As figured in Eq. 4, this similarity definition depends solely on the least common parent, therefore sim(People,Pets) = sim(Outfield Players,Bulldog) because they share the same least common parent: "Mammals".

$$\mathrm{sim}(c1, c2) = -\log \mathrm{p}\big(\mathrm{Iso}(c1, c2)\big) \quad (4)$$

Similar to Resnik, Lin [1] proposes another similarity definition based on the information content. It focuses on finding a definition that achieves two goals. The first goal is the *Universality*, by being able to use this definition in all probabilistic models (which are integrated in very different domains). The second goal is *Theoretical Justification.* The definition contains the same components as Resnik's measure, however the combination is not a difference, but it is a ratio as formulated in Eq. 5.

$$\mathrm{sim}(c1, c2) = \frac{2*\log \mathrm{p}\big(\mathrm{Iso}(c1,c2)\big)}{\log \mathrm{p}(c1)+\log \mathrm{p}(c2)} \quad (5)$$

In addition, Jiang and Conrath [4] calculates semantic distance in order to obtain the semantic similarity between two concepts, knowing that the semantic distance is the opposite of the semantic similarity. As formulated in Eq.6, the distance between two concepts is the difference between the sum of the information content of the two concepts and the information content of their most informative parent.

$$\mathrm{sim}(c1, c2) = 2 * \log \mathrm{p}\big(\mathrm{Iso}(c1, c2)\big) - \log \mathrm{p}(c1) + \log \mathrm{p}(c2) \quad (6)$$

3) Feature based measures:

Feature based measures, unlike the previous measures, are independent from the taxonomy and the parents of the concepts. These types of approaches are based on the assumption that each concept is described by a set of words that indicates their definitions and properties (features). And the more common features two concepts have, and the less non-common features they have, the more similar these two concepts are.

The Tversky measure [29] takes into account the features of terms to compute similarity between concepts, while ignoring the position and the content of those terms. This measure (Eq.7) considers that similarity is not symmetric, as features between a subclass and its superclass have a larger contribution to the similarity evaluation than those in the inverse direction.

$$\mathrm{sim}(c1, c2) = \frac{|C1 \cap C2|}{|C1 \cap C2| + \propto |C1 - C2| + (\propto -1)|C2 - C1|} \quad (7)$$

$\propto \in [0,1]$: The relative importance of the non-common characteristics (the value of $\propto$ increases with commonality and decreases with the difference between the two concepts).*

In addition, Petrakis et al. [30] proposes a feature-based function called X-similarity, which proposes a matching between words extracted by parsing their term definitions ("glosses" in WordNet or "scope notes" in MeSH). This measure is considered as a cross ontology semantic similarity, which means that it compares terms from different ontologies (WordNet and MeSH), but it can also be used for matching terms in the same ontology. Two terms are similar if the concepts of the words and the concepts in their neighborhoods are lexically similar, i.e. the more common words the definition of the concepts have, the more similar they are (and reversely).

4) Distributional based measures:

In this kind of approaches, similarity measure is based on the assumption that semantically close terms tend to appear in similar context.

This type of approach is used by Cilibrasi and Vitanyi [2], which proposes a similarity definition based on a background contents consisting of a documents database. The main idea of this proposal is to find similarity relations between two objects, by just using the number of documents (number of web pages in this case) in which the objects occur alone and together using a search engine (Google in this case). After obtaining the desired numbers, we can calculate the normalized Google distance (NGD) using Eq. 8, which gives a number between 0 and 1 to indicate the relationship between the two terms, knowing that 0 means that the two objects are identical and 1 means the non-existence of any relation between the two objects.

$$\mathrm{sim}(c1, c2) = \frac{\max\{\log \mathrm{f}(c1), \log \mathrm{f}(c2)\} - \log\ \mathrm{f}(c1,c2)}{\log \mathrm{N} - \min\{\log \mathrm{f}(c1), \log \mathrm{f}(c2)\}} \quad (8)$$

**f(c_i): The number of pages containing c_i.*
**f(c_i,c_j): The number of pages containing both c_i and c_j.*
**N: Number of Pages indexed by Google.*

Another similarity measure, proposed by Hindle [33] depending on mutual information estimated from text. The general idea in this approach is that "In any natural language there are restrictions on what words can appear together in the same construction. In particular, there are restrictions on what can be arguments of what predicates". Therefore, each noun will be characterized according to the verbs that it occurs with, then the nouns that appear the most in a similar context will be grouped together. For example, "Pizza", "Burger" and "Laptop" can all appear with the verbs "Sell" and "Buy", but only "Pizza" and "Burger" can appear with the verb "Eat". Which involves that "Pizza" and "Burger" are more similar to each other than they are similar to "Laptop".

5) Hybrid measures:

This type of approaches combines the ideas presented above in order to obtain a similarity definition with a higher accuracy than most of the basic edge-counting measures

An application of this type of measure in Pedersen's et al. approach [3], where they used the lexical database WordNet to improve the semantic measure concept by defining similarity using six already mentioned measures : Three measures based on path length between concepts (Path, Leacock & Chodorow, and Wu & Palmer) and three others based on information content (Resnik, Lin, and Jiang & Conrath) and defining relatedness using three measures. They showed their proposition using a software named WordNet::Similarity,

which takes the two terms as input and returns a number representing the degree of similarity or relatedness.

In addition, Knappe [34] defines a similarity measure between two concepts upon the set of all possible paths in the graph, using the information of generalization and specification as in Eq. 9.

$$\mathrm{sim}(c1,c2) = p * \frac{|\mathrm{Ans}(c1) \cap \mathrm{Ans}(c2)|}{|\mathrm{Ans}(c1)|} + (1-p) * \frac{|\mathrm{Ans}(c1) \cap \mathrm{Ans}(c2)|}{|\mathrm{Ans}(c2)|} \quad (9)$$

$p \in [0,1]$; The generalization influence degree
$Ans(c_i)$: The ancestor nodes of c_i.

Finally, Zhou et al. [35] proposes a hybrid similarity measure that takes into account information-based measures and path-based measures as parameters.(Eq. 10)

$$\mathrm{sim}(c1,c2) = 1 - k * \frac{\log(\mathrm{Length}(c1,c2)+1)}{\log\left(2*\left(\mathrm{Max}_{\mathrm{depth}}-1\right)\right)} - (1-k) * ((\mathrm{IC}(c1) + \mathrm{IC}(c2) - 2 * \mathrm{IC}(\mathrm{Iso}(c1,c2))/2) \quad (10)$$

k: Parameter that needs to be adapted manually for good performance. (If k=1 → Eq. 10: Path-based measure; If k=0 → Eq. 10: Information Content based measure)
IC: Information content

6) Discussion

Table II contains the comparison of the principles, the advantages and disadvantages of the presented type of Similarity Measures.

TABLE II. *Comparison of the different Semantic Measures Categories*

Category	Principle	Advantages	Disadvantages
Path Length based measures	Function of path length between concepts and their positions in the taxonomy	Simple to understand and implement	Two concepts with: 1- The same common parent. 2-Equal lengths of shortest path. will have the same similarity
Information Content based measures	The more common information two concepts share, the more similar these concepts are	1- Symmetric 2- Takes the information content of concepts into account	Two concepts with: 1- The same common parent 2- The same summation of information content will have the same similarity
Feature based measures	Concepts with more common features and less non common features are more similar	Takes concepts' features into account	Not effective when there is not a complete features set
Distributi-onal based measures	Similar concepts tend to appear in similar context	Works in all languages	1- Not effective with polysemic and empty words 2- Requires a large context
Hybrid measures	Combine multiple information sources	Distinguishes well different concepts pairs	1- Complex to implement 2- Needs to manually adapt parameters

B. Paraphrase Identification

In addition of proposing semantic measure definitions, paraphrase identification corresponds to the ability of determining the semantic similarity between two distinct strings. This is an important factor to improve cross-referencing methods for web information sources. However, we shall distinguish between the two terms *paraphrasing* and *textual entailment*. Two sentences or two longer natural language expressions are considered as a paraphrase for one another, if the two carry almost the same information. On the other hand, one sentence or one natural language expression is considered as a textual entailment of another, when by reading the first sentence you can infer that the second one is most likely also true. In fact, paraphrasing can be seen as bidirectional textual entailment, which leads that methods from the two areas are often similar For example, (A) and (B) are paraphrases, and each one textually entails (C), as knowing that the Eiffel Tower is visited by many tourists, a human can normally infer that Paris has many tourists.

(A) The Eiffel Tower is visited by thousands of tourists every year.
(B) Each year, thousands of people come to see the Eiffel Tower.
(C) We can find many tourists in Paris.

Many methods have been proposed that are related to paraphrasing and textual entailment. These methods can be classified into three categories [36]: Whether they perform paraphrase or textual entailment, *recognition, generation, or extraction.* The methods concerned about improving paraphrase generation are beyond the scope of this paper, as we are only interested in studying the methods that identify paraphrases, not the methods that generate them from a natural language expression.

1) Paraphrase and textual entailment recognition

These type of methods, judge whether or not a pair of natural language expressions are paraphrases or a correct textual entailment pair. The judgment is normally probabilistic, and serves to agree as much as possible with the judgment of humans.

The first set of approaches are *logic-based*, as they tend to represent the natural language expressions in a logical meaning, and then use theorem provers to check if these expressions are paraphrases or a pair of logical entailment. This kind of approaches requires the use of knowledge resources, like WordNet in Bos and Market's approach [37], Extended WordNet in Russ and Moldovan's approach [38], or

FrameNet in Baker's et al. approach [39]. An example of the usage of logical expressions in paraphrasing: A "pianist" is a hyponym (a specific case) of "musician" in WordNet. Therefore the following logical expression could be used to detect paraphrases or textual entailment in some cases:

$$\forall x \, \text{Pianist}(x) \Rightarrow \text{Musician}(x)$$

A different and a much less explored type of approach, that uses *vector space models of semantics* are employed to recognize paraphrases and textual entailment. This type of approach, used by Lin [40], maps each word of the two input language expressions to a vector that shows how strongly a word occurs with particular other words in a textual corpus. Then, the vectors of the single words can be combined into a compositional vector-based meaning representation. Furthermore, Pado and Lapata [41] takes into consideration the syntactic information of the expressions, as the meaning of a word or an expression can significantly change in a different syntax. For instance, even though (D) and (E) contain almost the same words, they are not paraphrases neither a pair of textual entailment expressions.

(D) They thought that Barcelona won the Spanish League.
(E) Barcelona won the Spanish League.

In addition, another type of approach based on *surface string similarity* are used to recognize paraphrases. This type of approach can use directly the two input language expressions, or can apply some pre-processing on them, like part-of-speech tagging or named entity recognition. The two input language expressions are considered as surface strings, and combinations of several string similarity measures are used to recognize paraphrases. For instance, Levenshtein [42] computed the string edit distance of the two input expressions, while Papineni et al. [43] and Zhou et al. [44] uses a well-known measure called BLEU. This measure examines the percentage of word n-grams (sequence of consecutive words) in two input strings and takes the geometric average of the percentages obtained for different values of n. As we can obviously see, (F) textually entails (G). However after using such n-gram measures, the similarity of the two expressions is low, as they have different lengths.

(F) The deadliest military conflict in history, World War II ended in 1945. Over 60 million people were killed, which was over 3% of the 1939 world population (est. 2 billion).
(G) World War II ended in 1945.

Several paraphrase and textual entailment recognition approaches operate at the *syntax level*. This type of approach, used by Melcuk [45] and Kubler et al. [46], employ dependency grammar parsers, that output a graph (usually a tree) whose nodes are the words of the expression and whose edges correspond to syntactic dependencies between words. For instance, we can see in Figure 2 [36] how the two expressions look very similar when viewed at the level of dependency trees, despite their differences in word order. We mention that these two expressions are not paraphrases, but (I) textually entails (H) because it contains the word "young".

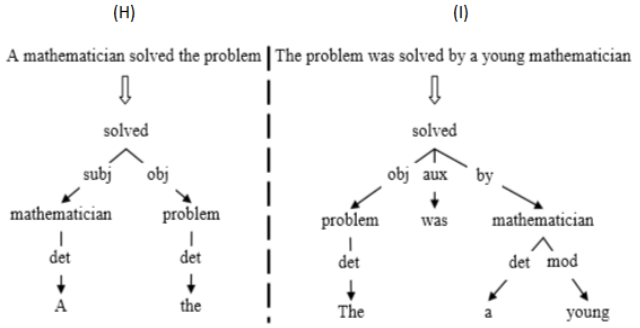

Figure 2. Two sentences viewed at the level of dependency trees

Another type of approach, recognizes paraphrases via *similarity measures operating on symbolic meaning representations*. This type of approach, used by Haghighi [47] and Marquez et al. [48], also uses graphs like the previous approaches. However, the edges of the graphs do not correspond to syntactic dependencies, but they reflect semantic relations figured in the input expressions. In order to compute these semantic relations, and as mentioned in Section II, different knowledge resources like FrameNet, PropBank and WordNet are used. This kind of approach is effective with paraphrases like (J) and (K) that have very similar meanings, with very few common words.

(J) The reputation of this man is going up.
(K) This guy's popularity is increasing.

2) Paraphrase and textual entailment extraction

Extraction methods, unlike recognition methods, does not classify input natural language expressions as paraphrases or not. However, their objective is to extract paraphrases or textual entailment pairs from large corpora. Paraphrase and textual entailment extraction are considered as a type of event extraction, which we discuss later in this paper.
The first type of extraction approach is based on the *distributional hypothesis*. For instance, Manning and Shuetze [49] proposes to represent by a vector all the word n-grams that occur in a large corpus with their left and right contexts. It considers as paraphrases the n-grams that occur frequently in similar contexts, by comparing their vectors using similarity measures like the ones discussed previously. In addition, Lin and Pantel [50] extends the distributional hypothesis type of approach, by operating it in their well-known method DIRT at the syntactic level.

In addition, Szpektor et al. [51] applies *bootstrapping* in their approach TEASE to extract paraphrases. TEASE starts with a lexicon of terms of a knowledge domain (e.g. names of symptoms and diseases in a case of a medical domain). These lexicons could well be automatically constructed from domain-specific corpus (medical articles in this case) using

term acquisition techniques. Then TEASE identifies noun phrases that co-occur frequently with each term of the lexicon, to use them as seed slot values in order to obtain templates. Finally, these templates are used in order to obtain more slot values. Figure 3 [36] shows an example of generating paraphrases by bootstrapping.

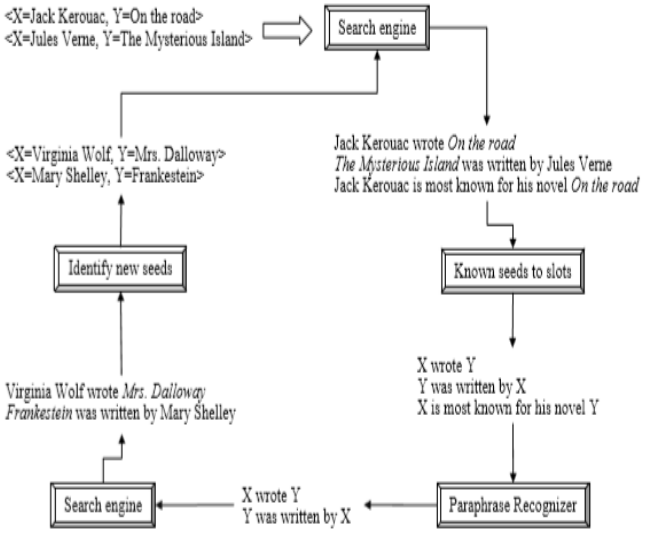

Figure 3. Generating paraphrases of "X wrote Y" by bootstrapping

Finally, Barzilay and Lee [52] proposes a paraphrase extraction method based on *multiple-sequence alignment.* Firstly, this approach starts by applying hierarchical complete-link clustering to sentences from two comparable corpora (different sources of the same gender, e.g. news articles from two press agencies). After obtaining clusters of sentences referring to events of the same type (event like "bomb attacks" in general, not a particular bomb attack), a word lattice was produced by aligning the cluster's sentences with multiple sequence alignment as shown in Figure 4 [36]. The two lattices <Slot 1> <bombed> <Slot 2> and <Slot 3> <was bombed by> <Slot 4> were drawn from different corpora. In the first lattice we have the sentence, which is represented by a path from start to end, "Planes bombed Baghdad". In the second lattice we have the sentence "Baghdad was bombed by planes". Finally, we deduce that the two lattices may well be paraphrases, where Slot 1 is assimilated with Slot 4, and Slot 2 is assimilated with Slot 3.

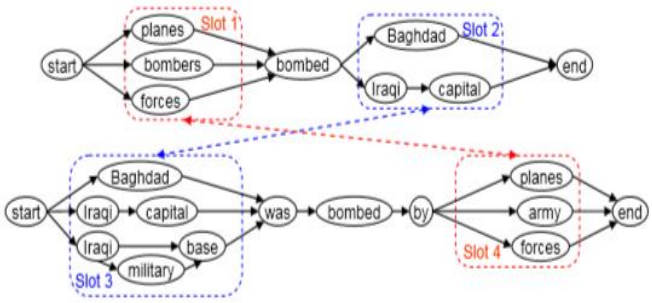

Figure 4. Word lattices obtained from sentence clusters

3) Discussion

In Table III, we present the tasks where the type of approaches discussed can be used. We mention that:
- TE: Textual Entailment; - P: Paraphrase; - Rec: Recognition; - Gen: Generation; - Ext: Extraction.

We can notice from this table, that all approaches used in paraphrase or textual entailment extraction, can also be used in their generation. In addition, we can see that both extraction and generation approaches are rarely used for paraphrase or textual entailment recognition and vice versa.

TABLE III. *Type of Approaches discussed and Tasks they have mostly been used in*

Type of Approaches	TE Rec	P Rec	TE Gen	P Gen	TE Ext	P Ext
Logic-based	×	×				
Vector space models of semantics		×				
Surface string similarity	×	×				
Syntax level	×	×				
Similarity measures operating on symbolic meaning representations	×	×				
Distributional hypothesis			×	×	×	×
Bootstrapping			×	×	×	×
Multiple-sequence alignment	×			×		×

C. Event Extraction

Finally, the last step to improve cross-referencing methods for web information sources is working on improving event extraction techniques. Event extraction, which is a common application of text mining, is the process of deriving high quality information from a certain text by identifying events. An example of a category of events is "sports matches" by considering the representation <Team> <Play Against> <Team>. A representation of this event category can be extracted from news headers such as "Manchester United hosted Liverpool", "Lakers will face Knicks", or "Real Madrid played against Barcelona".

However, as Hogenboom et al. [10] cite, we can distinguish between three main methods of event extraction: *data-driven* event extraction, *knowledge-driven* event extraction, and *hybrid* event extraction.

1) Data-Driven Event Extraction

This type of Event Extraction includes all approaches that

aim to convert data into knowledge. Data-driven approaches are based on word frequency counting, word sense disambiguation, N-grams and clustering, and rely only on quantitative methods to discover relations, such as the use of statistics, machine learning, linear algebra, information theory and probabilistic modeling. These approaches discover relations in corpora without considering semantics, and require a large amount of data to get statistically solid results, without the need of any linguistic resources or expert knowledge.

Several examples of the usage of this type of approach can be found in literature. For instance, Okamoto et al. [11] presents in 2009 a method for detecting occasional or volatile locale events such as events in restaurants or singers on street corners, by applying a two-level hierarchical clustering method using time-series blog entries collected with search queries. In addition, Liu et al. [12] in 2008, also uses clustering techniques after modeling entities and news documents as weighted undirected bipartite graph, in order to extract important events from daily web news. Finally, in 2005, Lei et al. [14] employs word-based statistical text mining, based on the use of subject extraction and an improved support vector machine in order to detect news and unreported real-life events.

2) Knowledge-Driven Event Extraction

Contrary to Data-driven methods, this type of event extraction includes all approaches that extract knowledge through representation and exploitation of expert knowledge. In order to extract information, these approaches use predefined or discovered linguistic patterns. These patterns can be either lexico-syntactic patterns or lexico-semantic patterns. In lexico-syntactic patterns, there are basically two important tasks to be performed. The first task is to identify the constituents (or sentence fragments) that a particular sentence is composed of. And the second task is to assign the exact roles to individual words, taking into consideration the grammatical rules of a language and a description (grammar) of how words can be put together in that language. For instance, taking the following sentence: "Most European countries, especially France, Germany, and Spain have good football teams". Applying a lexico-syntactic pattern that indicates the hyponymy (a type-of) relationship:

NP {,} especially {NP ,}{or | and} NP*

We can derive the following sentence fragments: hyponym ("France", "European Country"), hyponym ("Germany", "European Country"), and hyponym ("Spain", "European Country").However, at the lexico-semantic level, word disambiguation and occasionally the introduction of new alternative concepts in the representation of a text take place. Lexico-semantic procedures are normally used for augmenting syntactic analysis. Several attempts have been made for extracting events using mostly manually created lexico-syntactic patterns. For example, in their 2009 work, Nishihara et al. [15] proposes a system that uses lexico-syntactic patterns to split sentences in order to obtain personal experiences from blogs by visualizing an event as three kinds of pictures: action, object, and place.

A similar approach can be found in [16], where Aone et al. developed in 2000 an event extraction system using also lexico-syntactic patterns covering a wide range in the domains of finance and politics. This system consists of three specialized pattern-based tagging module, a high-precision co-reference resolution module, and a configurable template generation module.

In 2001, Yakushiji et al. [17] designs and implements a system to extract information in biomedical domain, based on full parsing with a large-scale, general-purpose grammar. Firstly the parser is used to convert text into canonical structure, and then a lexico-syntactic pattern is applied in order to extract information.

In the last example of the lexico-syntactic patterns employment, Xu et al [18], proposes in 2006 an approach to automatically detect events in prize-winning domain by learning patterns that signals the mentioning of such events. This approach is based on interpreting the event types as relations and starts with a set of seeds (which are semantic relations e.g. Subject "Company" – Verb "Appoint" – Object "Person").

In order to solve some of the limitations caused by using only lexico-syntactic patterns, different approaches have been proposed based on employing lexico-semantic patterns. For example, Li et al. [19] introduces a method to automatically discover event pattern in Chinese from stock market bulletin, based on the tagged corpus and the domain model. In addition, Cohen et al. [20] employs in 2009, a concept recognizer on a biological domain in order to extract medical events from corpora, taking into account the semantics of domain concepts.

And finally a similar approach is used by Vargas-Vera and Celjuska [21] in 2004, proposing a system that learns and apply lexico-semantic pattern in order to recognize events on news stories from Knowledge Media Institute (KMi) news articles.

3) Hybrid Event Extraction

In order to improve the results by benefiting from the advantages of Data-driven and Knowledge-driven event extraction approaches, several approaches were proposed based on combining techniques from these two approaches. Most hybrid systems are based on Knowledge-Driven approaches aided by Data-Driven methods, in order to solve the lack of expert knowledge or apply bootstrapping to boost extraction performances. For instance, in 2008, Tanev et al. [13] uses clustering methods to automatically tag words, in order to extract violent and disaster events from online news

and to automatically learn patterns from discovered events.

In addition, Cimiano and Staab [22] uses hybrid approaches in their software PANKOW to solve the lack of expert knowledge in pattern-based approaches. It was done by adding statistics methods using Google API to count the number of occurrences of a certain semantic or ontological relation.

Another use of Hybrid Event Extraction approaches in Jungermann and Morik's [23] approach in 2008, where they studied the case of extracting events from the minutes of plenary sessions of the German parliament. This case is part of developing a Question-Answering approach based on combining lexico-syntactic patterns with conditional random fields.

And finally, Chun et al. [24], develops a new unsupervised approach in order to extract events from biomedical literature based on using lexico-syntactic patterns in addition to term co-occurrences information and pattern score.

4) Discussion

From the results presented in Table IV, which provides a summary of the event extraction methods discussed, we derive that in terms of data usage, knowledge-driven event extraction methods require the least amount of data while data-driven methods require the most. On the other hand, we notice that data-driven methods require the least knowledge and expertise, while knowledge-driven methods require the most. And finally, we can derive that knowledge-driven methods and especially the methods that use lexico-semantic patterns are normally the easiest to be translated to a human-understandable format, while data-driven methods are the worst in this case. We can also conclude that hybrid event extraction methods show always a compromise between the two other methods, as this type of event extraction needs a medium amount of data usage (due to the use of data-driven methods), a medium amount of expert knowledge (lack of domain knowledge can be compensated by the use of statistical methods), and a medium level of interpretability (more difficult than knowledge-driven methods due to the addition of data-driven methods)

TABLE IV. An Overview of Event Extraction Methods

Method	***Data***	***Knowledge***	***Expertise***	***Interpretability***
Data	*High*	*Low*	*Low*	*Low*
Knowledge	*Low*	*High*	*High*	*High*
Hybrid	*Medium*	*Medium*	*High*	*Medium*

IV. DISCUSSION

The current paper discusses a large variety of approaches to natural language processing from diverse fields. When focusing on cross-referencing web information sources, one must instantly focus on extracting knowledge from these sources. But extracting knowledge from such dynamic and vast source is considered as the main challenge. However, we believe that using the features of the web, and considering them as characteristics instead of disadvantages will be the key. As we know, none of the knowledge resources mentioned will ever keep track with the web's development speed other than the web itself. As presented in Table I, we can always see the compromise between the size, the coverage, the structure and the growth of the knowledge resources. Nevertheless, we believe that improving the web's lack of structure will be the most efficient solution. In addition to choose the web as a start to extract knowledge, selecting the most suitable event extraction technique will be vital for any approaches' chance of success. Therefore, we suggest to benefit from the large data offered by the web, and consider *data-driven* approaches as the most suitable event extraction techniques. These approaches aim at converting data into knowledge relying on quantitative methods such as clustering and the use of statistics. However, event extraction techniques, as we mentioned in section I, depend on paraphrase identification methods to identify events expressed in different ways. Therefore, choosing paraphrase extraction approaches based on *distributional hypothesis*, and paraphrase recognition approaches based on *surface string similarity*, are a good solution as they both depend directly on semantic measures, which can be used to benefit from the presence of large context like the use of *distributional based measures.*

Since no approach is yet proved as the most efficient and reliable, one must choose, regarding the context of the issue, the most suitable combination of approaches. This choice depends in the first place on the knowledge resource used, then the event extraction technique. Finally, it depends on the best match between the paraphrase identification techniques and the similarity measure.

V. CONCLUSION

Cross-referencing web information sources is becoming one of the most important research fields. Its importance is due to the necessity of developing the existing web into a more intelligent and meaningful web. A web where machines are able to analyze and retrieve all of its data. However, this idea is not accomplished yet, which is why it is expected to see many other approaches in the near future. This paper is considered as a start on integrating knowledge in the web, by providing an advanced examination of cross-referencing methods of web information sources. In order to examine these methods, we started by investigating the most recognized semantic measures that can be used to evaluate the resemblance between concepts or groups of concepts. Then, we studied approaches on paraphrase and textual entailment identification. Finally, we investigated approaches to event extraction from text. They can be clustered into three types: data-driven, knowledge-driven and hybrid event extraction methods. In addition to cross-referencing web information sources, these approaches can be used in several other natural language processing tasks including question answering, text

summarization, text generation, and machine translation applications.

ACKNOWLEDGMENT

We wish to thank the “Conseil Régional de Bourgogne” for its help in supporting and funding this work.

REFERENCES

[1] Dekang Lin, An Information-Theoretic Definition of Similarity. Department of Computer Science University of Manitoba Winnipeg, Manitoba, Canada R3T 2N2

[2] Rudi L. Cilibrasi and Paul M.B. Vitanyi, The Google Similarity Distance. IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No 3, March 2007, 370–383

[3] Ted Pedersen, Siddarth Patwardhan and Jason Michelizzi, WordNet::Similarity - Measuring the Relatedness of Concepts, Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04), pp. 38-41, May 3-5, 2004, Boston, MA. (Demonstration System)

[4] Jay J. Jiang and David W. Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. Proceedings of International Conference Research on Computational Linguistics (ROCLING X), 1997, Taiwan

[5] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney, Impact of Similarity Measures on Web-page Clustering. The University of Texas at Austin, Austin, TX, 78712-1084,USA

[6] Chris Brockett and William B. Dolan, Support Vector Machines for Paraphrase Identification and Corpus Construction. Natural Language Processing Group Microsoft Research One Microsoft Way, Redmond, WA 98502, U.S.A

[7] Vasile Rus, Philip M. McCarthy, Mihai C. Lintean, Danielle S. McNamara and Arthur C. Graesser, Paraphrase Identification with Lexico-Syntactic Graph Subsumption. University of Memphis Departments of Computer Science, Psychology, and English Institute for Intelligent Systems Memphis, TN 38152, USA

[8] Philip Resnik, Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Department of Linguistics and Institute for Advanced Computer Studies, University of Maryland, Journal of Artificial Intelligence Research 11 (1999), 95 -130

[9] Yoan Chabot and Christophe Nicolle, Semantic Measures: A State of the Art. The Encyclopedia of Information Science and Technology 10, 2014

[10] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong: An Overview of Event Extraction from Text. Erasmus University Rotterdam

[11] Okamoto, M., Kikuchi, M.: Discovering Volatile Events in Your Neighborhood: Local-Area Topic Extraction from Blog Entries. In: 5th Asia Information Retrieval Symposium (AIRS 2009). Lecture Notes in Computer Science, vol. 5839, pp. 181– 192. Springer-Verlag Berlin Heidelberg (2009)

[12] Liu, M., Liu, Y., Xiang, L., Chen, X., Yang, Q.: Extracting Key Entities and Significant Events from Online Daily News. In: 9th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2008). Lecture Notes in Computer Science, vol. 5326, pp. 201–209. Springer-Verlag Berlin Heidelberg (2008)

[13] Tanev, H., Piskorski, J., Atkinson, M.: Real-Time News Event Extraction for Global Crisis Monitoring. In: 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems (NLDB 2008). Lecture Notes in Computer Science, vol. 5039, pp. 207–218. Springer-Verlag Berlin Heidelberg (2008)

[14] Lei, Z., Wu, L.D., Zhang, Y., Liu, Y.C.: A System for Detecting and Tracking Internet News Event. In: 6th Pacific-Rim Conference on Multimedia (PCM 2005). Lecture Notes in Computer Science, vol. 3767, pp. 754–764. Springer-Verlag Berlin Heidelberg (2005)

[15] Nishihara, Y., Sato, K., Sunayama, W.: Event Extraction and Visualization for Obtaining Personal Experiences from Blogs. In: Symposium on Human Interface 2009 on Human Interface and the Management of Information. Information and Interaction. Part II. Lecture Notes in Computer Science, vol. 5618, pp. 315–324. Springer-Verlag Berlin Heidelberg (2009)

[16] Aone, C., Ramos-Santacruz, M.: REES: A Large-Scale Relation and Event Ex- traction System. In: 6th Applied Natural Language Processing Conference (ANLP 2000). pp. 76–83. Association for Computational Linguistics (2000)

[17] Yakushiji, A., Tateisi, Y., Miyao, Y.: Event Extraction from Biomedical Papers using a Full Parser. In: 6th Pacific Symposium on Biocomputing (PSB 2001). pp. 408–419 (2001)

[18] Xu, F., Uszkoreit, H., Li, H.: Automatic Event and Relation Detection with Seeds of Varying Complexity. In: AAAI Workshop on Event Extraction and Synthesis (2006)

[19] Li, F., Sheng, H., Zhang, D.: Event Pattern Discovery from the Stock Market Bulletin. In: 5th International Conference on Discovery Science (DS 2002). Lecture Notes in Computer Science, vol. 2534, pp. 35–49. Springer-Verlag Berlin Heidelberg (2002)

[20] Cohen, K.B., Verspoor, K., Johnson, H.L., Roeder, C., Ogren, P.V., Baumgart- ner, Jr., W.A., White, E., Tipney, H., Hunter, L.: High-Precision Biological Event Extraction with a Concept Recognizer. In: Workshop on BioNLP: Shared Task collocated with the NAACL-HLT 2009 Meeting. pp. 50–58. Association for Computational Linguistics (2009)

[21] Vargas-Vera, M., Celjuska, D.: Event Recognition on News Stories and Semi- Automatic Population of an Ontology. In: 3rd IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004). pp. 615–618 (2004)

[22] Cimiano, P., Staab, S.: Learning by Googling. SIGKDD Explorations Newsletter 6(2), 24–33 (2004)

[23] Jungermann, F., Morik, K.: Enhanced Services for Targeted Information Retrieval by Event Extraction and Data Mining. In: 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems (NLDB 2008). Lecture Notes in Computer Science, vol. 5039, pp. 335–336. Springer-Verlag Berlin Heidelberg (2008)

[24] Chun, H.W., Hwang, Y.S., Rim, H.C.: Unsupervised Event Extraction from Biomedical Literature Using Co-occurrence Information and Basic Patterns. In: 1st International Joint Conference on Natural Language Processing (IJCNLP 2004). Lecture Notes in Computer Science, vol. 3248, pp. 777–786. Springer-Verlag Berlin Heidelberg (2004)

[25] Lingling Meng, Runqing Huang and Junzhong Gu, A Review of Semantic Similarity Measures in WordNet. In: International Journal of Hybrid Information Technology,Vol. 6, No. 1, January, 2013

[26] C. Leacock and M. Chodorow, “Combining Local Context and WordNet Similarity for Word Sense Identification, WordNet: An Electronic Lexical Database”, MIT Press, (1998), pp. 265-283.

[27] Z. Wu and M. Palmer, “Verb semantics and lexical selection”, Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, (1994) June 27-30; Las Cruces, New Mexico

[28] Slimani, T.: Description and Evaluation of Semantic Similarity Measures Approaches. Computer Science Department Taif University & LARODEC Lab

[29] Tversky, A. 1977. Features of Similarity. Psycological Review, 84(4):327-352

[30] Petrakis, E. G. M., Varelas, G., Hliaoutakis, A., & Raftopoulou, P. 2006. X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies. Journal of Digital Information Management, 4, 233-237.

[31] Steve Lawrence and C. Lee Giles. "Accessibility and Distribution of Information on the Web," Nature 400(6740): 107-109, July 8, 1999

[32] "Alexa Top 500 Global Sites". Alexa Internet. Retrieved 8 December 2014.

[33] Hindle, D. (1990). Noun classification from predicate argument structures. Proceedings of the 28th annual meeting on Association for Computational Linguistics (pp. 268-275).

[34] Knappe, R., Bulskov, H. and Andreasen, T. 2003. On Similarity Measures for Content-Based Querying. In O. Kaynak, editor, Proceedings of the 10th International Fuzzy Systems Association World Congress (IFSA'03), pages 400-403, Istanbul, Turkey, 29 June - 2 July.

[35] Z. Zhou, Y. Wang and J. Gu, "New Model of Semantic Similarity Measuring in WordNet", Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, (2008) November 17-19, Xiamen, China.

[36] Androutsopoulos, Ion and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research, 38:135-187.

[37] Bos, J., & Markert, K. (2005). Recognising textual entailment with logical inference. In Proc. of the Conf. on HLT and EMNLP, pp. 628–635, Vancouver, BC, Canada.

[38] Moldovan, D., & Rus, V. (2001). Logic form transformation of WordNet and its applicability to question answering. In Proc. of the 39th Annual Meeting of ACL, pp. 402–409, Toulouse, France.

[39] Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. In Proc. of the 17th Int. Conf. on Comp. Linguistics, pp. 86–90, Montreal, Quebec, Canada.

[40] Lin, D.(1998b). An information-theoretic definition of similarity. In Proc. of the 15th Int. Conf. on Machine Learning, pp. 296–304, Madison, WI. Morgan Kaufmann, San Francisco, CA.

[41] Pad´o, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. Comp. Ling., 33(2), 161–199.

[42] Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physice-Doklady, 10, 707–710

[43] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J.(2002). BLEU: a method for automatic evaluation of machine translation. In Proc. Of the 40th Annual Meeting on ACL, pp. 311–318, Philadelphia, PA.

[44] Zhou, L., Lin, C.-Y., & Hovy, E. (2006a). Re-evaluating machine translation results with paraphrase support. In Proc. of the Conf. on EMNLP, pp. 77–84.

[45] Melcuk, I. (1987). Dependency Syntax: Theory and Practice. State University of New York Press.

[46] Kubler, S., McDonald, R., & Nivre, J. (2009). Dependency Parsing. Synthesis Lectures on HLT. Morgan and Claypool Publishers

[47] Haghighi, A. D. (2005). Robust textual inference via graph matching. In Proc. of the Conf. on EMNLP, pp. 387–394, Vancouver, BC, Canada.

[48] M`arquez, L., Carreras, X., Litkowski, K. C., & Stevenson, S. (2008). Semantic role labeling: an introduction to the special issue. Comp. Linguistics, 34(2), 145–159.

[49] Manning, C. D., & Schuetze, H. (1999). Foundations of Statistical Natural Language Processing. MIT press.

[50] Lin, D., & Pantel, P. (2001). Discovery of inference rules for question answering. Nat. Lang. Engineering, 7, 343–360.

[51] Szpektor, I., Tanev, H., Dagan, I., & Coppola, B. (2004). Scaling Web-based acquisition of entailment relations. In Proc. of the Conf. on EMNLP, Barcelona, Spain.

[52] Barzilay, R., & Lee, L. (2003). Learning to paraphrase: an unsupervised approach using multiple sequence alignment. In Proc. of the HLT Conf. of NAACL, pp. 16–23, Edmonton, Canada.