Context and Discourse in Textual Entailment Inference

Shachar Mirkin

Department of Computer Science

Ph.D. Thesis

Submitted to the Senate of Bar-Ilan University

Ramat Gan, Israel

October 2011

This work was carried out under the supervision of Prof. Ido Dagan (Department of Computer Science), Bar-Ilan University.

Acknowledgements

Many people have helped making the period of my Ph.D. studies successful and enjoyable and here I wish to express my gratitude to them. First and foremost, I would like to thank my advisor, Prof. Ido Dagan for his guidance and support during my graduate studies. Ido has been a role model and a source of inspiration in matters way beyond the scope of my research, and I feel truly fortunate for having him as my advisor. I am grateful to the co-authors of the papers included in this thesis: Jonathan Berant, Lili Kotlerman, Sebastian Padó, Eyal Shnarch, Idan Szpektor, Nicola Cancedda, Marc Dymetman and Lucia Specia. These collaborations have been both educating and a pleasure. To the latter three, Nicola Cancedda, Marc Dymetman and Lucia Specia, as well as to Wilker Aziz, I also wish to thank for their great hospitality during my visits to XRCE and for the collaboration on the joint project. I thank my colleagues and friends at the NLP lab at Bar-Ilan University: Naomi Frankel, Asher Stern, Eyal Shnarch, Chaya Liebeskind, Amnon Lotan, Chen Erez, Ephi Sachs, Hadas Zohar, Libby Barak, Roni Ben-Aharon, Roy Bar-Haim, Iddo Greental, Erel Segal, Meni Adler, Hila Weiss, Idan Szpektor, Jonathan Berant and Lili Kotlerman. I'm glad I had the opportunity to be part of such a unique and diverse group of people who made my time there more fruitful, interesting and fun. I am thankful to the staff of the secretary office of the Department of Computer Science, Dafna Gad, Sylvie Baruch, Hila Atia and Sigal Segev-Zerahia, for their continuous assistance and for making my life much easier when dealing with bureaucracy.

My family, Inbal, Ofir, Gil and Nitzan, as well as my parents, are really the ones who made it all possible for me through their endless love and support. I am deeply grateful and indebted to them. Lastly, I wish to extend my gratitude to all the people who drove their cars in the Aluf Sade interchange that day in March 2007 when I decided to go for a Ph.D. The massive traffic jam they created led me through a detour right by the university. Things would have been entirely different for me if it wasn't for them.

Preface

Portions of this thesis are joint work and have appeared elsewhere.

Chapter 3, "Evaluating the Inferential Utility of Lexical-Semantic Resources", appeared in the Proceedings of the 12$^{\text{th}}$ Conference of the European Chapter of the Association for Computational Linguistics (EACL-09) (Mirkin et al., 2009b). Chapter 4, "Source-Language Entailment Modeling for Translating Unknown Terms", appeared in the Proceedings of the Joint Conference of the 47$^{\text{th}}$ Annual Meeting of the Association for Computational Linguistics and the 4$^{\text{th}}$ International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP-2009) (Mirkin et al., 2009c). Chapter 5, "Classification-based Contextual Preferences", appeared in the Proceedings of the 2011 Workshop on Applied Textual Inference (TextInfer), EMNLP 2011 (Mirkin et al., 2011). Chapter 6, "Assessing the Role of Discourse References in Entailment Inference", appeared in the Proceedings of the Conference of the 48$^{\text{th}}$ Annual Meeting of the Association for Computational Linguistics (ACL-2010) (Mirkin et al., 2010b). Chapter 7, "Recognising Entailment within Discourse", appeared in the Proceedings of the 23$^{\text{rd}}$ International Conference on Computational Linguistics (Coling 2010) (Mirkin et al., 2010a).

This work was partially supported by the Negev Consortium of the Israeli Ministry of Industry, Trade and Labor, the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886 and the Israel Science Foundation grants 1095/05 and 1112/08.

Abstract

Semantic inference is key to Natural Language Processing. A Question Answering system which is requested to answer the question "*When was Hong Kong returned to China?*" can naively do so by searching for instances of "*Hong Kong was returned to China in ...* ". It may thus be able to extract the answer from instances like "*Hong Kong was returned to China in 1997*", but will fail to do so from texts such as "*In the summer of 1997, the eyes of the world were on Hong Kong as China assumed control over the former crown colony.*" Being able to conclude that this text contains the answer being sought is the kind of task for which semantic inference over texts is required. Such inference, also referred to as *textual inference*, is thus the task of drawing conclusions from textual assertions and in particular inferring one textual assertion from another, even when the two are phrased differently.

Human language exhibits various phenomena which make inference a challenging task. One such phenomenon is lexical ambiguity. For instance, the word *assumed* means '*took*' in the above sentence, but may mean '*presumed*' in other contexts. Another phenomenon is manifested in the fact that not every piece of information explicitly appears in the text. Many times information is omitted or implicitly-referred to, relying on readers' ability to fill in the gaps. While inference often depends on reasoning and world knowledge that go beyond the given text, much of the non-explicit information can be recovered from the text itself. In the above example, inferring

that China assumed control over Hong Kong is based on understanding that *the former crown colony* refers to *Hong Kong*. Both these phenomena represent a common property of human language: words are interpreted based on the context in which they occur. In this thesis we investigate these two contextual aspects of language: (i) the impact of the context on word meaning, which particularly concerns ambiguous words, and (ii) the way pieces of information that are scattered across multiple locations of the text are jointly used to determine the interpretation of words. Through theoretical and empirical studies we wish to shed light on the ways in which these phenomena affect inference and to identify directions for better utilizing them in order to improve semantic inference.

In recent years, Textual Entailment (TE) has become a popular paradigm for modeling semantic inference. The core TE task, Textual Entailment recognition, is to determine whether the meaning of one text can be inferred – or *entailed* – from another. As such, Textual Entailment provides a common methodology for addressing inference tasks across text understanding applications. We thus adopt the Textual Entailment framework in this thesis for investigating the impact of context and discourse phenomena on inference.

As demonstrated in the above example, computational semantic inference relies on the availability of knowledge that would make up for world and linguistic knowledge which is naturally available to humans. Such knowledge is often found within repositories of semantic relationships, and can be made available in the form of *inference rules* (or *entailment rules*). Typically, an inference rule consists of two text fragments, one of which is expected to entail the other (e.g. '*China* $\Rightarrow$ *country*'). While many knowledge resources are available, only a few of them are comprised of entailment relations, and even fewer were evaluated with respect to inference needs. For

that reason we designed an evaluation methodology for assessing the practical contribution of lexical-semantic resources to inference and applied it to several distinct resources. Our study reveals that employing knowledge in inappropriate contexts, a consequence of language ambiguity, constitutes the primary contributor to erroneous inferences and is therefore the major cause for degrading resources utility (Chapter 3; Mirkin et al., 2009b).

Contextual validation of rule application was carried out in our next work, where we took up the task of translating unknown terms in Machine Translation through an entailment-based approach. During translation, Machine Translation systems frequently encounter unknown (out-of-vocabulary) terms. To address this problem, we used monolingual source-language data, namely inference rules, to directly modify source sentences using terms which are known to the translation system. This process generates paraphrases of a source sentence or sentences that are entailed by it. The usage of inference rules is contextually validated in order to ensure that only context-appropriate modifications are applied. For instance, if the word *exercised* in the source sentence "*This strategy was exercised in the 1980s*" is unknown to the translation system, it may be replaced by known words, such as *used*, once the context of *exercised* in the source sentence is verified to match the replacement. This method significantly improves the coverage of the translation system, as judged according to the number of accepted translations, and outperforms previous methods suggested for this task (Chapter 4; Mirkin et al., 2009c).

In the above Machine Translation work, we looked at a single contextual aspect of inference – the contextual match of the given text to the inference rule. Yet, other types of *context matching* should be considered. For example, in an Information Extraction task, we may wish to ensure that only rules which contextually match the target event are used for extraction. E.g., for a *release* event, which refers to 'end of custody', we need to discard rules concerning software releases. We thus

suggest a generic classification-based scheme which handles context matching in a comprehensive and coherent way, allows the integration of various types of contextual information and captures the notion of directionality within entailment inference. Our scheme, implemented and applied on a Text Categorization task, outperformed the state of the art on this task (Chapter 5; Mirkin et al., 2011).

Interpretation of words in context concerns more than just ambiguous words. References between words, their position in the text and information conveyed elsewhere in the text are among the factors that play a role in the way words are understood by the readers. All these fall into the category of *discourse*, which in our work refers to the entire provided text and accordingly includes information beyond the local environment of the considered words. Discourse-induced references between words have long been considered by text understanding applications through the use of coreference resolvers. Typically, anaphors were substituted by their antecedents or mapped to them, thus making the referent explicit and easier to interpret. This limited use of discourse information undermines the potential impact of discourse on inference. We thus set out to investigate what other types of discourse information may be useful for textual inference and how inference systems can use such information. Among our findings, we discovered that references involving verbal phrases and bridging relations are significant for inference, and that inference systems should incorporate other operations than just the ones used today in order to consider discourse information (Chapter 6; Mirkin et al., 2010b).

We further proceeded to empirically explore the potential of discourse-induced information to practically improve inference systems' performance. We developed methods for utilizing various types of discourse information within entailment systems. Our results significantly surpassed those of an identical system differing only in the way it attended to discourse, and outperformed state of the art entailment systems (Chapter 7; Mirkin et al., 2010a).

In sum, although it is common knowledge that contextual information can be beneficial for inference, it has so far been considered in quite limited ways. Altogether, our work shows that various types of information embedded in the context – in its many aspects and scopes – are significant for inference and should be given priority and demonstrates that utilizing contextual information can practically lead to better inference performance. We have displayed such improvements through a number of novel algorithms, experiments and analyses, suggesting several research directions for utilizing the context for further enhancing inference systems and the modules which they use.

Outline

This thesis is comprised of five published works. Generally speaking, these works can be mapped into three different sub-tasks of Textual Entailment systems: (i) the use of lexical entailment knowledge, (ii) context matching and (iii) discourse processing. Chapter 3 addresses the first task, Chapters 4 and 5 the second and the third is covered by Chapters 6 and 7.

More specifically, the thesis is organized as follows.

Chapter 1 provides an introduction to semantic inference in general and to Textual Entailment in particular. Sections 1.3, 1.4 and 1.5 present each of the three tasks we address and introduce our approach to each one of them.

Chapter 2 describes the methodologies which we used in order to assess the novel methods proposed in this work, as well as the evaluation and the analysis methodologies that we suggest for assessing the contribution of certain components to inference. Chapter 3 presents an evaluation methodology for assessing the practical utility of lexical-semantic resources when used as sources for entailment rules. This methodology is applied to several such resources, providing comparable indicators of their performance and insights about the cases where they provide inappropriate rules. Chapter 4 proposes a novel method to address unknown terms in Machine Translation without relying on multilingual corpora. Unknown terms in the source text are replaced by entailing ones, while the context-appropriateness of the substitution is assessed through context models. While in Chapter 4 we make use of known

context models and address a specific aspect of context matching, in Chapter 5 we suggest a novel scheme for comprehensively addressing context matching in inference scenarios and demonstrate the scheme using a Text Categorization application. Chapters 6 and 7 address the impact of discourse on inference and show, analytically (Chapter 6) and empirically (Chapter 7), the contribution of discourse information to inference performance. Chapter 8 briefly summarizes the insights and contributions of this work and suggests directions for future research.

Contents

List of Abbreviations

Abbreviation	Meaning
BIUTEE	Bar-Ilan University Textual Entailment Engine
CP	Contextual Preferences
IE	Information Extraction
IR	Information Retrieval
LHS	Left-hand side
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MT	Machine Translation
NB	Naïve Bayes
NLP	Natural Language Processing
POS	Part of Speech
QA	Question Answering
RHS	Right-hand side
RTE	Recognising Textual Entailment
SMT	Statistical Machine Translation
SP	Selectional Preferences
SVM	Support Vector Machine
TC	Text Categorization
TE	Textual Entailment
WSD	Word Sense Disambiguation

Chapter 1

Introduction

1.1 Semantic inference

Semantic inference is the task of making inferences over natural language texts based on their meaning. Such inferences are prevalent in Natural Language Processing (NLP) applications. For instance, in Information Retrieval (IR), it is necessary to infer whether a document satisfies the information need which is conveyed by a natural language query; in Information Extraction (IE), a potential occurrence of an event in the text should imply the meaning of the event; and in Question Answering(QA), an answer is expected to be inferred from candidate retrieved passages.

The task is, naturally, very challenging. One prominent reason is language variability: the meaning of every textual assertion can be expressed in multiple ways. For instance, the two sentences in Example 1.1 convey the same meaning.

(1.1) (a) *By 2022, Germany plans to abandon nuclear energy.*

(b) *Germany decides to drop nuclear power by 2022.*

Such texts are referred to as *paraphrases*. In NLP applications, it is desirable

to be able to identify paraphrasing texts. For example, in Multiple-document Summarization (henceforth 'Summarization'), identifying paraphrases is useful in order to avoid including redundant information in the summary. Therefore, a summary about Germany's decision should include either sentence 1.1a or sentence 1.1b, but not both.

Moreover, a sentence can be discarded from being included in a summary even if it is not a paraphrase of any sentence already in the summary; rather, it is sufficient to identify that the meaning of the new sentence is implied, or *entailed*, from the summary. Consider Example 1.2. In this example, sentence 1.2b is implied from 1.2a, thus a summary that contains 1.2a need not include 1.2b as well.

(1.2) (a) *German parliament votes for closure of all nuclear plants by 2022.*

(b) *Germany plans to shut down nuclear power plants.*

It turns out that capturing this kind of inferential relation between texts is a common task which is shared amongst multiple text understanding applications. *Textual Entailment* captures exactly this idea.

1.2 Textual Entailment

Textual Entailment (Dagan and Glickman, 2004; Dagan et al., 2009) is a generic paradigm for applied semantic inference which captures inference needs of many NLP applications. Under Textual Entailment, applications' inferences are reduced to the common task of determining whether the meaning of a given textual assertion, termed *hypothesis* (H), can be inferred from the meaning of a certain *text* (T). If the answer is positive, we say that T *(textually) entails* H, and denote the relation '$T \Rightarrow H$'.[1]

[1]Note that the term 'Textual Entailment' is used to denote both the inference paradigm and the relation between texts.

Example 1.2 is an instance of a (textual) entailment relation, where 1.2a entails 1.2b. In this example, sentence 1.2b takes the hypothesis role and 1.2a the text's.

Textual Entailment is a directional relation. It is easy to observe that sentence 1.2b does not entail 1.2a, since 1.2a contains information that is missing from 1.2b. When the entailment is bi-directional, i.e. when both texts mutually entail each other, we denote the relationship '$T \Leftrightarrow H$'. Under this perspective, therefore, paraphrasing is actually a specific case of the Textual Entailment relation.

The usefulness of identifying entailment relationships for Summarization was described above. Another example is Text Simplification (Inui et al., 2003). In this application, the task is to generate a simpler version of an input text. The simplified text may be useful, for example, for providing easy-to-read texts for low-literacy readers (Aluisio et al., 2010). The relation between the original text and its simplified version corresponds to entailment, where the original text paraphrases or directionally entails the simplified one. Example 1.3 demonstrates the entailment relationship in a Text Simplification scenario. Sentence 1.3a is the first sentence from the Wikipedia English definition of *cloud*[1], while 1.3b is the first sentence from the *cloud* definition in Simple English Wikipedia[2]. Here, 1.3a entails 1.3b.

(1.3) (a) *A cloud is a visible mass of water droplets or frozen ice crystals suspended in the atmosphere above the surface of the Earth or another planetary body.*

(b) *A cloud is water in the atmosphere (air) that we can see.*

Modeling Textual Entailment has been proven useful for many other applications including Information Extraction (Romano et al., 2006), Question Answering (Harabagiu and Hickl, 2006; Iftene and Balahur-Dobrescu, 2007; Negri et al., 2008),

[1] http://en.wikipedia.org/wiki/Cloud, August 2011
[2] http://simple.wikipedia.org/wiki/Cloud, August 2011

Machine Translation Evaluation (Padó et al., 2009), Summarization (Harabagiu et al., 2007) and Intelligent Tutoring (Nielsen et al., 2009). In Chapter 4 we show how Textual Entailment modeling is used to improve Machine Translation (MT) performance, and in Chapter 5 we demonstrate how a Text Categorization (TC) task is mapped into Textual Entailment.

In comparison with the notion of entailment in formal semantics (Chierchia and McConnell-Ginet, 2000), the Textual Entailment relation is not expected to be satisfied in each and every case. In other words, the inference needs not be unquestionably certain. It is rather a somewhat relaxed relation which is assumed *true* if "a human reading T would infer that H is most likely true" (Dagan and Glickman, 2005). It is therefore better-suited for applied settings, as the relation will not be considered *false* even if there is an anecdotal situation in which the relation does not hold. Suppose, for example, that we wish to generate a summary about Germany's withdrawal from using nuclear energy. For most practical purposes, an inclusion of sentence 1.2a in the summary makes sentence 1.2b redundant, even though it is possible to find circumstances where this is not entirely accurate. Such an unlikely situation may occur if the German parliament's decision is somehow overruled. Yet, it is reasonable to believe that most human readers will not consider such an option viable and will practically regard the meaning of 1.2b as implied from 1.2a. The Textual Entailment definition follows this rationale.

Entailment and generalization

In many cases (yet not always), a directional entailment relationship coincides with the notion of "generalization", where H conveys a more general meaning of T. This frequently comes into play when H refers to more general concepts or contains fewer details than T. The idea is demonstrated in the following example through details that are absent in H relative to T, and through the use of a hypernym in H (*a run*)

instead of its hyponym in T (*the London Marathon*).

(1.4) T : *In 2002, Lloyd Scott completed the London Marathon wearing a deep sea diving suit that weighed a total of 50KG.*

H : *Lloyd Scott participated in a run wearing a diving suit.*

This is the basis for many inference operations that are being employed in Textual Entailment models. The use of hypernymy or member-set relationships, the removal of redundant adjectives or even of entire clauses from T are all examples for such generalizations. This is also the underlying idea of our approach in Chapter 4, where a more general (entailed) sentence is translated by a Machine Translation system when the exact meaning of the source sentence cannot be properly translated.

Nevertheless, this is but one aspect of entailment relations between texts. In Example 1.4 above, the understanding that Mr. Scott participated in the run stems from the knowledge expressed in T that he *completed* the run. This is by no means a generalization, but rather a deduction based on world-knowledge that also falls under Textual Entailment, as do various other semantic relations.

1.2.1 Recognising Textual Entailment (RTE)

The standard Textual Entailment modeling task is termed Textual Entailment *Recognition*. Given a text T and a hypothesis H, the recognition task is to determine whether T entails H, i.e. to recognise instances of entailment. Various techniques have been proposed to tackle this task. Approaches branch out as early as the choice of text representation and preprocessing. While some systems remain at the raw textual level (Castilloa and i Alemany, 2008; Galanis and Malakasiotis, 2008), other represent texts and hypotheses as dependency parse trees (Bar-Haim et al., 2009; Padó et al., 2008) or even through logical representations (Bos and Markert, 2005; Clark and Harrison, 2008).

Textual Entailment inference is usually comprised of a set or a sequence of inferences, based on matches between the text and the hypothesis. Such matches may be *direct*, when the same text fragment appears in both T and H (as the name *Lloyd Scott* in Example 1.4), or *indirect*, relying on semantic relationships between terms in the texts, as between *Marathon* and *run*. One popular Textual Entailment modeling approach constructs an alignment, or a mapping, between entities in T and H, such as between words or parse tree edges, and determines entailment by the overall quality of the alignment (Padó et al., 2008; Zanzotto et al., 2009; Burchardt et al., 2009; MacCartney et al., 2008). Another approach is based on transformations, where T is explicitly modified by a series of knowledge-driven entailment transformations such that it becomes more similar to H. Each transformation preserves the entailment relation, i.e. the consequent text is entailed by the text from which it was generated (Bar-Haim et al., 2008a; Harmeling, 2009). Other methods include the translation of T and H into logical representations and applying theorem provers to determine whether logical entailment holds (Bos and Markert, 2005), or measuring tree edit distance between T and H (Kouylekov and Magnini, 2005; Cabrio et al., 2008).[1]

Regardless of representation and the inference approach, a common goal is shared by many Textual Entailment systems: obtaining maximal *coverage* of the hypothesis' components by the text.[2] Such hypothesis components may include words (e.g. Clark and Harrison, 2010), parse tree nodes or sub-trees (Bar-Haim et al., 2008b), named entities (Castilloa and i Alemany, 2008), events (Wang and Neumann, 2008) or other types of textual entities. Coverage is used to determine entailment through various techniques. A popular one is classification features which quantify the coverage by T of specific H components, such as verbs, nouns or named entities (Bar-Haim et al., 2009). Alternatively, the degree of covered H components in an aligned model can

[1] See (Androutsopoulos and Malakasiotis, 2009) for a more complete review of Textual Entailment modeling approaches.

[2] Since Textual Entailment is a directional relation, the coverage sought is also non-symmetric.

be used as a threshold to decide whether T entails H (Marsi et al., 2006).

1.2.2 Entailment knowledge and entailment rules

The entailment recognition process typically depends on knowledge. This knowledge may specify that a *boxer* is a *dog*, that *Manhattan* is located in *New York* and that if a certain drug *cures* a disease it means that the drug *treats* the disease.

One common way to encode such knowledge is through *inference rules*, which in the context of Textual Entailment are commonly referred to as *entailment rules*. The first two among the above knowledge items can be encoded with the following entailment rules:

(1.5) (a) '*boxer* $\Rightarrow$ *dog*'

(b) '*Manhattan* $\Rightarrow$ *New York*'

As with the Textual Entailment relation, we use the notation '$\Rightarrow$' to indicate that the meaning of the left-hand side (LHS) of the rule entails the meaning of its right-hand side (RHS).[1] Such knowledge is useful in applied inference settings. For instance, when searching for "*Kosher Indian restaurants in New York*", an answer can be obtained from the sentence "*A new Indian glatt kosher restaurant opened in Manhattan, on 63rd and 1st.*" based, among other things, on the rule '*Manhattan* $\Rightarrow$ *New York*'.

When the rule contains only lexicalized items (i.e. actual words) we refer to it as a *lexical entailment rule*. Rules may also include *templates*, i.e. expressions with variables. Such rules typically correspond to entailment between predicate-argument constructs, e.g. '*X cure Y* $\Rightarrow$ *X treat Y*'. Typically, such rules are specified along with the syntactic relation between their elements. For instance '$X \xleftarrow{subj} cure \xrightarrow{obj} Y$',

[1] See Section 1.3 regarding *lexical entailment* – entailment at the lexical level.

indicates that X is the subject of *cure* and Y is its object. Consequently, such rules are often called *lexical-syntactic rules*. Using templates allows more precise inferences to be made, as syntactic structures are being matched, and since restrictions may be assigned to the variables. Such restrictions usually take the form of Selectional Preferences (SP), which specify the types of entities that may instantiate the templates' variables. In the above example, X may be instantiated with types of drugs, e.g. *antibiotics* or with specific drug names, e.g. *Doxycycline*, while Y may be instantiated with types or names of diseases, e.g. *sinus infection*.

Syntactic rules make a third kind of entailment rules. These rules represent a transformation between two syntactic constructs, which are comprised mostly of variables, function words and the syntactic relations between them. For instance, using dependency parse tree representation, the following rule may be employed to transform a parse tree from passive form to active form.

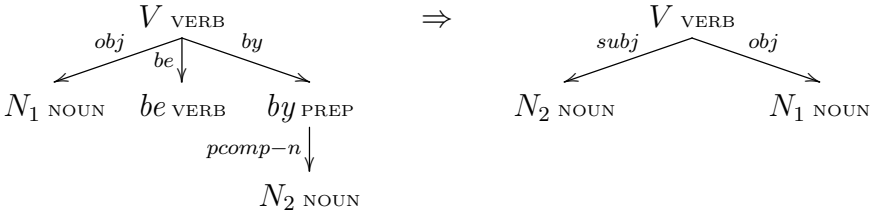

Figure 1.1: Passive to active entailment rule as in (Bar-Haim et al., 2007). N_1, N_2 and V are the rule's variables. Relation labels in this figure are based on the MINIPAR dependency parser (Lin, 1998c).

Entailment rules are being used in virtually all entailment systems. In alignment-based entailment systems, the rule and its semantic type enable the alignment of non-identical nodes and determine the link's score (e.g. alignment between synonyms is preferred over other non-identical alignment types); in transformation-based systems, rules induce the modifications of the text; and in systems which determine entailment based on edit distance between T and H, distances are computed as a direct result

of the relations encoded by the employed rules (Kouylekov and Magnini, 2006).

Rule application

The usage of inference rules within the entailment process, in particular in transformation-based systems, is often called *rule application*. A rule is applicable when its LHS is *matched* in T, where the goal is to match the RHS in H in order to infer the matched RHS from the LHS. The match in H can be either immediate, through the application of a single rule, or following a sequence of rule applications, where the RHS of the last applied rule is matched in H. Lexical rules are matched when the LHS is found in the text, while lexical-syntactic and syntactic rules require a structural (syntactic) match as well. Depending on the Textual Entailment modeling approach and the task at hand, a rule application may be explicit, resulting in the actual substitution of the LHS in T by the RHS, or implicit, by matching a component in T with a component in H, as in alignment-based Textual Entailment systems.

Sources of entailment rules

Lexical-semantic relations have been the focus of numerous studies. Relations were either manually specified, (e.g. WordNet; Fellbaum, 1998) or automatically acquired using methods based on patterns (Hearst, 1992; Brin, 1998; Berland and Charniak, 1999; Chklovski and Pantel, 2004; Snow et al., 2004), distributional similarity (Lin, 1998a; Lindn and Piitulainen, 2004; Yates and Etzioni, 2009a; Kotlerman et al., 2010; Schoenmackers et al., 2010) and combinations thereof (Mirkin et al., 2006; Pennacchiotti and Pantel, 2009), as well as a collection of other methods including probabilistic models (Snow et al., 2006) or graph walks (Hughes and Ramage, 2007).

Many researchers have made the output of their works available to the community in the form of semantic resources. Clearly, the most commonly used resource among them is WordNet, which contains a variety of semantic relations between nouns, verbs,

adjectives and adverbs, including synonymy, hyponymy and antonymy, to name a few. Other resources, e.g. Rion Snow's Augmented Wordnets (Snow et al., 2006), provide *is-a* relationships much like WordNet's hyponymy relationships, but mainly focus on named entities rather than generic words, thus complementing WordNet's rather sparse *hyponym-instance* relation. Other types of lexical relations represented within available resources include similarity (Lin, 1998a; Chklovski and Pantel, 2004), derivational-relatedness, (e.g. Nomlex (Macleod et al., 1998) and CatVar (Habash and Dorr, 2003)) and "enablement" (e.g. *fight–win* in VerbOcean (Chklovski and Pantel, 2004)). Additional released resources are not purely lexical but rather capture relations between templates. Such resources contain both paraphrases[1] (Lin and Pantel, 2001; Sekine, 2005; Yates and Etzioni, 2009b) and directional inference rules (Szpektor et al., 2004; Schoenmackers et al., 2010).

Yet, until recently, only a few available resources were dedicated to entailment relationships. Consequent knowledge gaps still constitute a major obstacle for entailment systems and inference-based applications (Bos and Markert, 2005; Balahur et al., 2008; Bar-Haim et al., 2008c). In the absence of sufficient knowledge, systems resort to heuristics and approximation methods, rather than strive to achieve a complete and established entailment proof.

Sometimes, acquiring entailment relationships from such resources is straightforward. WordNet's hyponyms, for example, typically entail their hypernyms, and clearly synonyms correspond to (bi-directional) entailment as well. Similarly, synonyms can be extracted from the members of a verb class in VerbNet (Kipper et al., 2000). On the other hand, many resources provide relations that do not coincide exactly with entailment. Similarity-based resources, for instance, may include synonymy and hyponymy relationships, but also non-entailing antonyms and cohyponyms (a.k.a.

[1]The term 'paraphrases' is used in the literature to describe any set of texts or text fragments with equivalent meanings, from single words through predicate-argument tuples to entire sentences. In the context of inference rules, typically the shorter fragments are referred to.

coordinate terms). Lexical-syntactic resources for paraphrases supposedly provide bi-directional entailment relationships. Yet, in practice, such resources correspond to a more fuzzy notion of paraphrasing than necessary in inference tasks. The reason is that such resources, apart from being noisy as are all automatically-generated resources, contain both bi-directional and directional relationships, albeit without specifying the direction. For instance, one popular set of resources is based on the DIRT algorithm (Lin and Pantel, 2001). These resources, automatically constructed based on corpus statistics, contain both proper paraphrases (in the sense of mutual entailment), such as '*X search for Y* $\Leftrightarrow$ *X look for Y*', but also '**X search for Y* $\Leftrightarrow$ *Y found by X*'[1], which is not symmetric.

This is the case with many other resources that are being exploited for entailment modeling purposes, and are therefore problematic when used for inference tasks. Still, there is more to the entailment relation than synonymy and is-a relations, and knowledgebases such as Dekang Lin's distributional similarity resources were successfully employed for RTE (Marsi et al., 2006; Padó et al., 2008) and other inference-based tasks (Zhang et al., 2004; Padó et al., 2009; Surdeanu et al., 2011).

1.3 The impact of lexical entailment resources

As mentioned above, a multitude of lexical-semantic resources are available, yet lexical rules used for entailment are commonly derived from resources which were not designed specifically for entailment purposes. Consequently, reported evaluations of such resources are largely uninformative from an entailment perspective. It is therefore necessary to understand the actual contribution of such knowledgebases to inference when used as sources of entailment rules. To that end, in Chapter 3 we present a methodology for the evaluation of the practical utility of lexical-semantic resources

[1]DIRT rules are lexical-syntactic – the syntactic dependencies are removed here for readability; asterisks mark incorrect entailment rules.

and apply it to seven prominent resources. The idea of our evaluation methodology is to narrow the inference down to a single operation – a lexical rule application in a given context – and judge whether the operation resulted with a correct inference.

Sub-sentential entailment and lexical inferences

Within Textual Entailment inference, lexical-level inferences are very common. Lexical relations currently constitute the most widely used type of knowledge for entailment tasks, due to their independence of the text representation, their wide availability and their (relatively) high coverage. In fact, purely-lexical Textual Entailment systems (i.e. those that use only lexical rules) have thus far proven competitive in comparison with complex RTE systems (Adams et al., 2007; Roth and Sammons, 2007; Mirkin et al., 2009a; Shnarch et al., 2011a,b).

It should be noted that the Textual Entailment relation is defined in terms of truth values: entailment between T and H holds if the truth of T implies the truth of H (Dagan and Glickman, 2005). While suitable for propositions – sentential assertions – this definition is not appropriate for sub-sentential text fragments and in particular for single words or terms, which are exactly the ones we are interested in for our evaluation. Hence, to assess the impact of a given lexical-semantic resource, a definition of entailment at the lexical level is required. Dagan and Glickman (2004) addressed sub-sentential (non-propositional) hypotheses by relating the concepts of entailment and existence, such that an expression is entailed from a sentential text if its existence is implied. Zhitomirsky-Geffet and Dagan (2009) defined entailment between two lexical expressions, but only for substitutable cases. For many purposes, such as Text Categorization (see Chapter 5), their definition is too restrictive, as many useful inferences, e.g. '*hospital* $\Rightarrow$ *medical*', are not substitutable. Glickman et al. (2006) defined an entailment relation between sentential texts and sub-sentential hypotheses, which they termed *lexical reference*. For our evaluation and analysis we

slightly adapt their definition and suggest in addition an operational definition for lexical entailment relations, i.e. relations between two sub-sentential text elements (see Chapter 3).

Instance-based vs. rule-based evaluation

One way to evaluate the accuracy of a knowledge resource is by manually judging the validity of a sample of its relationships. Annotators are shown a pair of textual expressions and judge whether they satisfy a certain relation (e.g. Snow et al., 2006). In the context of Textual Entailment, these evaluations are known as *rule-based*. An obvious drawback of such evaluations is that they are detached from actual applicative settings. Rules included in a resource, even if they correctly correspond to the specified relation, may be rarely required or highly ambiguous. In turn, this may result in infrequent or incorrect usage. An alternative evaluation approach is called *instance-based*, where judges are given actual instances on which the rules are being applied and need to decide whether their usage yielded the desired outcome. This is the kind of evaluation we conduct in our work.

The results of our comparative evaluation show that the manually-constructed WordNet ranks high in terms of precision also under instance-based evaluation. However, its coverage is limited and needs to be complemented by automatically-generated and consequently less accurate (from an entailment perspective) resources, such as Dekang Lin's word similarity resources (Lin, 1998b). This outcome stresses the need for further development of large scale resources for entailment rules.

One of the most prominent outcomes of our study is that the utility of even the most accurate resources is compromised when context is not taken into account. For instance, the rules in Example 1.5 ('*boxer* $\Rightarrow$ *dog*' and '*Manhattan* $\Rightarrow$ *New York*') include ambiguous terms. *Boxer* may refer to a boxing athlete or to a member of

the Chinese 1900 Boxer Rebellion, and *Manhattan* may refer to a Woody Allen film or to a cocktail drink. The rule '*Manhattan* ⇒ *New York*' will therefore result with an incorrect inference if applied to the following sentence: "*I really need a drink. I'll have a Manhattan, please.*" Hence, before a rule is applied, its appropriateness to the given context needs to be validated. We further elaborate about this issue in Section 1.4.

We note that in recent years, after the publication of the abovementioned work (Mirkin et al., 2009b), several entailment resources – both lexical and lexical-syntactic – have been made available by the Bar-Ilan University NLP lab[1]. These include resources based on Wikipedia (Shnarch et al., 2009) and FrameNet (Ben-Aharon et al., 2010), a directional distributional similarity resource (Kotlerman et al., 2010) and a knowledgebase of predicative entailment rules (Berant et al., 2011). The practical contribution of these resources to inference has yet to be analyzed. Some positive indications are provided in the above publications and in ablation tests conducted as part of recent RTE challenges (Bentivogli et al., 2009, 2010).

1.4 Context-sensitive inference

Lexical ambiguity poses an obstacle for many text understanding applications; consequently, determining word meaning in context is a fundamental NLP task. Over the years it has been shown that addressing ambiguity helps the performance of various NLP applications, including Information Retrieval (Stokoe et al., 2003), Question Answering (Harabagiu et al., 2003), Machine Translation (Carpuat and Wu, 2007) and Information Extraction (Szpektor et al., 2008). As mentioned in Section 1.3, this is a prominent task for inference, in particular for rule application, and if ignored it can undermine the practical usefulness of entailment knowledge resources.

[1] Available for download at http://www.cs.biu.ac.il/~nlp/downloads/

Word Sense Disambiguation

An active thread of research for coping with lexical ambiguity is *Word Sense Disambiguation* (WSD; Navigli, 2009). The WSD task is to choose the appropriate sense of a target word in context from among a predefined set of senses for the word, as specified within some sense inventory. For English WSD, the most popular sense inventory is WordNet. A known problem when using WordNet's senses is granularity. WordNet senses are often very fine-grained or refer to anecdotal meanings which makes the distinction between senses of a given word difficult even for humans. Further, the level of granularity needed is often task-dependent and language-dependent, making a single predefined set of senses less useful. Consequently, manual effort may be required to generate such inventories per language, task or domain of interest.

The OntoNotes project (Hovy et al., 2006) represents efforts to tackle the sense granularity problem. Taking a top-down approach, they first split word meanings into coarse-grained senses and gradually split them further into finer-grained ones. Throughout this process, 90% human agreement in sense annotation is maintained in order to avoid a too fine-grained distinction between senses. This method solves part of the abovementioned problems, while the issues of task- and language-dependence remain.

Direct Sense Matching

Direct sense matching is an alternative approach to WSD, applicable in inferential settings. It avoids explicitly determining the senses of the target words; instead, given a pair of word occurrences, the task is reduced to the binary question of determining only whether the meanings of the considered words do or do not *match* in the given context. That, without explicitly specifying each word's sense (Dagan et al., 2006a). Consider Example 1.6. The direct sense matching task is to determine whether the

sense of the word *disc* in 1.6b or 1.6c matches the sense of the word *record* in 1.6a. In this case, we expect the matching algorithm to answer "yes" for the sense match between *record* and *disc* in 1.6a and 1.6b, respectively, and "no" for the match between the senses of the two words in 1.6a and 1.6c.

(1.6) (a) *"On the last **record** we were consciously writing love songs. So we didn't want to repeat that."*

(b) *Celebrating the release of their latest **disc**, Bon Jovi will be performing live in Times Square.*

(c) *A prolapsed **disc** often causes severe lower back pain.*

Following this approach, obstacles arising from the use of sense inventories are removed. Such an approach is also frequently employed for *lexical substitution* tasks (McCarthy and Navigli, 2009), where the goal is to identify contextually-suitable substitutions for a target word occurrence.

An example of the utilization of this approach is provided in Chapter 4, where we apply lexical substitutions to generate paraphrases or entailed texts of sentences to be translated. To ensure that rules are applied in appropriate contexts, context models based on a language model, Naïve Bayes classification or Latent Semantic Analysis (LSA) are employed. In neither case do we explicitly identify the senses of the words in the text, but instead rank them according to scores of the context match between the text to be translated and the rule.

Context Matching

Dagan et al. (2006a) addressed sense matching of synonyms in order to determine their substitutability in a given context. A generalization of this approach is provided by the *context matching* task, the goal of which is to identify compatible contexts for text expressions which are not necessarily substitutable, and are not necessarily

lexical. Expressions need not be of the same part of speech, and may be comprised of syntactic elements as well. Such an approach is useful for applications where meaning correspondence is sought rather than substitution, such as Text Categorization or Information Retrieval. For instance, in Text Categorization it may be necessary to ensure that the meaning of the term *alien* in a document matches the meaning of the category *outer space*, although these terms are typically not substitutable. This is the task we undertake in Chapter 5 to ensure that inferences are context-sensitive.

Contextual Preferences

Considering actual inference scenarios, it turns out that three context matching aspects should be regarded between each pair among the participating *inference objects*: the (term in the) text, t, (the term in) the hypothesis h and the rule r. *Contextual Preferences* (CP; Szpektor et al., 2008) is a generic framework for inferential context matching that considers all of these aspects, thus providing the most comprehensive view to date of context matching in semantic inference. In comparison, prior work only addressed specific matches between these objects (Harabagiu et al., 2003; Pantel et al., 2007; Barak et al., 2009). Adhering to Textual Entailment, CP conforms with entailment directionality, stating a direction for each context match. This is a fundamental aspect of Textual Entailment that is largely overlooked. Suppose we wish to assess whether the Textual Entailment relation holds between each of T or T' and H in Example 1.7. One of the rules at our disposal is '*boxer* $\Rightarrow$ *dog*', but to employ it we first need to ensure that the contexts in which the rule is applied are valid for it. Since the goal is to infer H from T or T', and not vice versa, the context of the more specific word *boxer* in T or T' should match the meaning of the more general word *dog* as in H, and not the other way around (i.e., we do not need the context of *dog* in H to specifically match the meaning of *boxer* as in T or T'). In this case, *boxer* in T, but not in T', contextually matches *dog* in H.

(1.7) T : *Ear cropping is considered the breed standard for boxers in the US, while in some countries this procedure is outlawed.*

T' : *In a 1997 rematch in Las Vegas, boxer Mike Tyson was disqualified for biting off a piece of Evander Holyfield's ear, and his license was revoked.*

H : *Cropping a dog's ears is not prohibited in the United States.*

Based on the above considerations, when addressing context matching in Chapter 5, we adopt CP as our context matching framework.

In the CP framework, each of the three inference objects, t, h and r has a contextual representation, termed "CP". Before carrying out an inference operation, the CPs of the involved objects must be matched. Szpektor et al. (2008) suggested looking at two types of contextual information for each match: (i) the selectional preferences of variables in a template, which they called CP_v, and (ii) the topical or global context, denoted CP_g. Being a generic framework, any concrete context matching model can be applied within CP to address the matches between each pair among t, h and r. Szpektor et al. (2008) proposed different methods to address each of the two types (CP_g and CP_v) for each of the three context matches ($t-h$, $t-r$ and $r-h$). In addition, their proposed models were symmetric, thus practically ignoring the directionality aspect of their framework.

Our goal is to implement CP using a model which is unified for all context matching aspects, which captures directionality and which is capable of incorporating any type of contextual information. To achieve these properties, in Chapter 5 we propose a novel classification-based scheme for Contextual Preferences, which is comprised of two types of classifiers for addressing the three contextual matches specified by the CP framework. Contrary to the previous implementation of the CP framework, our scheme is directional, taking into account the different roles of the inference objects

in terms of context matching. Classifiers provide a convenient way for incorporating various types of contextual information, through features. Thus, the distinction between global context and variable instantiation CPs in the original framework is no longer necessary.

Classification-based context models

Classification has been used in prior work to address context matching. In classification-based models, typical valid and invalid contexts for a given inference object are automatically obtained and used to train a classifier. In turn, this classifier is used to assess the validity of a given new context for the object. For example, a classifier for the rule '*wing* ⇒ *group*', which refers to the political sense of *wing*, can be trained using texts such as Example 1.8a as positive examples, and texts like Examples 1.8b and 1.8c as negative examples.

(1.8) (a) *Uruguay's ruling coalition of left* ***wing*** *parties has a good chance to claim victory in this Sunday's ballot.*

(b) *A baby falcon is being treated for an injury to its left* ***wing*** *after the bird tried to leave her nest atop the federal courthouse.*

(c) *The left* ***wing*** *of the building houses the Stephen Foster Memorial Museum.*

Given a new text in which *wing* occurs, the classifier has to determine whether it constitutes a valid context for the application of the above rule (that is, whether it constitutes a positive or negative instance with respect to the trained rule classifier).

In prior classification-based models, only specific aspects of the possible contextual matches have been addressed, typically between t and r. Kauchak and Barzilay (2006) used context classifiers for Machine Translation Evaluation. In this task the

goal is to automatically evaluate the output of an MT system with respect to human-generated reference translations. Standard evaluation metrics such as BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) compute n-gram overlaps between the output and the references. Such metrics do not capture language variability as even simple semantic modifications are penalized. To circumvent this problem, Kauchak and Barzilay paraphrased reference sentences through synonym substitution, making them more similar to the MT output in their wording. To ensure that the substitution is contextually valid, they learned a classifier for each substituting word using occurrences of the word in a corpus as positive examples and random sentences as negative examples. Then, given the MT output sentence with the word to substitute, the context is validated by the substituting word's classifier before the actual substitution takes place. Dagan et al. (2006a) addressed synonym substitution in context by training a single-class SVM (Schölkopf et al., 2001), i.e. a classifier that uses only positive training examples and learns the smallest hypersphere which bounds most of them. They formulated lexical substitution as a query expansion task, where single-word queries are expanded with their synonyms (e.g. *disc* acts as the expansion of the query "*record*"). A classifier is trained for each query word, where corpus occurrences of the word are provided as positive examples from which standard WSD features are extracted. Given an occurrence of an expansion in the text (e.g. a sentence containing the word *disc*), its context is examined by the query classifier in order to determine whether it constitutes a valid match. Contexts are classified as appropriate for the query word if they fall inside its classifier's learned hypersphere.

The above works, like our own Machine Translation work (Chapter 4), address only the context match between t and r[1]. In addition, they only target substitutable

[1]Note that in this simple setting, much like Kauchak and Barzilay's, it is equivalent to view the context match as either a $t - h$ match, where t is the occurrence of the expansion word in the text and h is the query word, or as a $t-r$ match, where the LHS of the rule is the expansion and the query

words, synonyms in particular. Rather, it is desired to handle any entailment rules, whether substitutable or not, and to address all three CP aspects. Lastly, both Kauchak and Barzilay and Dagan et al. learned a classifier for each word using all of the senses of the word intermixed, which is suitable for monosemous words or biased towards the most common senses of a polysemous word. In our model (Chapter 5), we aim to learn the classifier for the specific intended meaning of the inference object.

Connor and Roth (2007) addressed verb paraphrasing in context by using a set of per-rule "local" classifiers to generate training data for a global classifier, using as many as hundreds of thousands of training examples. The global classifier's positive training data consisted of examples which the local classifiers were able to label confidently, while negative examples were based on randomly selected occurrences of verbs that are not semantically related to the verb being paraphrased. Six features were extracted from these examples, three of which considered context-insensitive overlap measures between the verb to paraphrase and its candidate substitution and the other three considered the overlap of the two verbs with respect to the given context. The main advantage of this work is that the global classifier can be applied to rules for which no classifier was trained in advance. Still, a set of occurrences of the verb-pair needs to be collected on-the-fly in order to apply the classifier. Like the two abovementioned works, this work addresses only the match between the text and the rule (t and r). In contrast, the scheme we propose addresses all of the context matching aspects specified under the CP framework.

1.5 Inference within discourse

A contextual aspect of language that plays an important role in textual inference is the impact of the *discourse*, i.e. the role of information conveyed in the entire

word is its RHS (e.g. '*disc* ⇒ *record*'). Within our work, we adopt the second perspective, which regards the context matching of the substitution as a validation of an entailment rule application.

text in determining the interpretation of each term and sentence in it. Processing discourse enables the integration of information that is scattered in multiple locations in the text into a unified, preferably explicit, representation. In this work the term 'discourse' generally refers to the scope of the text beyond a single sentence, but may also concern multiple text expressions within a single sentence, as demonstrated below.

Probably the most commonly known discourse phenomenon is *coreference*, a reference of multiple text expressions to the same entity or event. In Example 1.9, *Li Na* and *she* are related via a coreference relation as both refer to the Chinese tennis player, Li Na. Likewise, a coreference relates *China* and *the country* in this text.

(1.9) ***Li Na***$_{[1]}$ *was hugely popular in* ***China***$_{[2]}$ *long before* ***she***$_{[1]}$ *won the French Open final, becoming* ***the country***$_{[2]}$*'s first tennis player to lift a Grand Slam singles title.*

Example 1.9 presents two coreference relations between nouns. Nominal coreference relations are possibly the most common type of discourse-induced references, and are clearly the most studied one. Yet, verbal phrases also participate in coreference. Example 1.10 shows such a coreference relation, where *demonstrated* and *marched* refer to the same event. Information Extraction is a prominent application in which event coreference has been shown to be beneficial (e.g. Humphreys et al., 1997; Bejan and Harabagiu, 2010).

(1.10) (a) *In what are described as the largest rallies in Israel's history, hundreds of thousands have* ***demonstrated***$_{[1]}$ *against high living costs and demanded economic reform.*

(b) *Some 450,000 Israelis* ***marched***$_{[1]}$ *at massive rallies across the country.*

In addition to coreference relations, another, indirect, type of discourse reference is realized in Example 1.9, relating *Grand Slam* with *French Open*. This relation

clarifies, even for those who do not follow tennis regularly, that the French Open is a Grand Slam tournament. These and many other types of discourse-induced relations are often categorized as *bridging* references (Clark, 1975). Among the types of bridging relations that have been described in the literature are: parts of objects, events or situations (e.g. between *window* and *room* in *I walked into the room. The window looked out to the bay.*), set-membership, roles in events (e.g. *John was murdered yesterday. The murderer got away.*), reasons, causes and consequences (Clark, 1975; Asher and Lascarides, 1998). The automatic identification of bridging relations has received relatively little attention, and when it did, the focus was typically given to specific relations among them, namely meronymy (part-of), that could be identified via extraction patterns or that appear in WordNet (Markert and Nissim, 2003; Poesio et al., 2004).

To our knowledge, to date, available tools do not support the resolution of verbal coreferences and bridging relations. With respect to the latter, apart from being more difficult to annotate and to automatically identify, it remains generally unknown whether resolving bridging relations actually helps NLP applications.

These are among the reasons that most entailment systems have addressed discourse references only by the substitution of nominal coreference anaphors by their antecedents. One of the few exceptions is (de Marneffe et al., 2008), where text-hypothesis coreferring events were detected for the purpose of identifying contradicting texts. Yet, even in this work, event coreference was identified via text-hypothesis alignment rather than based on the output of a coreference resolution system.

In addition to coreference and bridging relations, other discourse-related phenomena have been studied, including lexical chains, coherence and topic continuity. Various discourse analysis techniques have been employed for applications such as Essay Analysis, Summarization and Question Answering (Burstein et al., 2003; Chai

and Jin, 2004; Clarke and Lapata, 2010).[1] Like discourse references, these have made almost no imprint on Textual Entailment systems.

In Chapter 6 we describe an analysis of the potential utility of discourse references (coreference and bridging) for inference. Particularly, we show that apart from nominal coreference relationships, verbal reference relations are also significant for inference, that bridging references are frequently realized in inferential scenarios and that operations other than substitution need to be applied by inference systems in order to take advantage of the information provided by the discourse. We also discuss the interplay between inference knowledge and discourse references, where both can provide supporting evidence to the relation between entities in the text, as in the case of *Grand Slam* and *French Open* in Example 1.9 above. In Chapter 7 we empirically address these and other discourse-related phenomena, showing that discourse-induced information can indeed be useful in improving the performance of Textual Entailment systems.

[1]For a more complete overview see the ACL 2010 Discourse Structure tutorial (Webber et al., 2010).

Chapter 2

Evaluation and Analysis Methodologies

This chapter presents the different evaluation and analysis methodologies that were used to evaluate each of the tasks we took up in this work. It follows two main approaches for assessing the performance of inference systems or their internal modules: an evaluation based on the performance of an end-application and a direct evaluation based on a gold-standard annotation of inference relations. The following two sections describe our work from that perspective. Section 2.3 briefly reviews the analysis methodology we suggested for assessing the impact of discourse references on inference.

2.1 Evaluation via application performance

Perhaps the ultimate way to assess the impact of a system's component is to show how it affects the system's performance on an end-application. At the end of the day, we are interested to know whether the component has a positive effect on the application as a whole, whether it degrades performance, or whether its impact is minimal.

In Chapter 5 we present a novel scheme for addressing contextual aspects in inference, as well as a concrete model implementing the scheme. An evaluation is conducted by measuring the performance of a Name-based Text Categorization algorithm when using our approach and when using other methods. Performance is measured with respect to a gold-standard annotation provided with the Text Categorization datasets.

In Chapter 4 we propose tackling unknown terms in Machine Translation by translating known entailed terms in their stead. For example, if *chalet* is unknown, we may translate *house* instead. The performance of our method is assessed on an English to French Machine Translation task, using both automatic and manual evaluations. For the automatic evaluation we used the standard MT Evaluation metrics BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007). In the manual evaluation, human annotators judged whether the translation was acceptable, or compared between the translations of two competing systems. Manual evaluation was required since automatic MT Evaluation metrics are typically based on sequence comparison between the translation of the MT system and human-generated reference translations. Such metrics, therefore, do not account well for semantic modifications of the type captured by Textual Entailment. Indeed, METEOR and metrics that followed its rationale made the first steps towards addressing this problem, but so far they have been limited to specific semantic relations, such as synonyms, and depend on resources like WordNet that are not available for every language.

2.2 Direct evaluation of inference relations

Evaluation based on an end-application is indeed desirable but is not suitable in all circumstances. One risk of conducting an application-based evaluation is that the complexity of the application, its setting, or the system in which a component is

embedded will mask the actual contribution of the assessed component. As a consequence, we may gain little knowledge regarding the actual weaknesses and strengths of the component or its performance relative to other methods for the same sub-task. In addition, available evaluation datasets are not always well-suited for the assessment of certain components. For example, even a perfect coreference resolution system will have little impact on the RTE-2 (Bar-Haim et al., 2006) dataset, since coreference issues were largely avoided in that dataset (see more details in Section 2.2.1). Therefore, in two of our works we conducted an evaluation in direct accordance with manual judgement of the entailment relationships between texts. In one case, the entire text is considered; in another, the evaluation zooms in on a single inference operation.

2.2.1 The Recognising Textual Entailment challenges

The Recognising Textual Entailment challenges (Dagan et al., 2009) became the main benchmark for assessing the performance of Textual Entailment systems. Since 2005, six annual sessions have been completed (Dagan et al., 2006b; Bar-Haim et al., 2006; Giampiccolo et al., 2007, 2008; Bentivogli et al., 2009, 2010).

The goal of the challenges is to evaluate the capability of RTE systems to correctly identify entailment relations. The traditional RTE task was, given a pair of texts, T and H, to determine whether T entails H. Texts and hypotheses were derived from data of various applications, including Information Extraction, Information Retrieval, Question Answering, Machine Translation Evaluation and Summarization. For instance, in RTE-2, $T - H$ pairs were generated from Question Answering data based on the TREC-QA[1] and QA@CLEF[2] datasets. Hypotheses were manually constructed from questions that were transformed into affirmative statements with the answers manually inserted into them; texts consisted of actual passages retrieved by

[1] http://trec.nist.gov/
[2] http://clef-qa.itc.it/

QA systems for the corresponding questions.

In the first two challenges texts were comprised of single sentences, while in RTE-3 (Giampiccolo et al., 2007) paragraph-length texts were introduced, thus making the issue of coreference resolution much more relevant. Still, texts were guaranteed to be self-contained, thus artificially reducing the role of discourse in the dataset. This situation has changed with the introduction of the RTE-5 Search task (Bentivogli et al., 2009). In that task, given a hypothesis H and a small corpus, the goal is to find all corpus sentences (i.e., Ts) that entail H. This time, texts were not modified and were left in their natural environments – the documents. A variant of the Search task became the main task in RTE-6 (Bentivogli et al., 2010) and in RTE-7[1]. Sentences were judged individually, but their interpretation considered the entire discourse (e.g. a mentioned location is assumed known in consequent sentences). In consequence, discourse information became much more substantial for entailment recognition. In all challenges, gold standard entailment judgements were manually generated.

In Chapter 7 we present a discourse-enhanced system built on top of the Bar-Ilan University Textual Entailment Engine (BIUTEE; Bar-Haim et al., 2008c, 2009). BIUTEE, which also acts as a baseline, addresses discourse only minimally, much like other Textual Entailment systems. In our evaluation over the RTE-5 Search task, our system outperforms the baseline system in terms of micro-averaged and macro-averaged Precision, Recall and F_1, and achieves a performance surpassing previously reported results for the task.

2.2.2 Evaluating single inference operations

In the work presented in Chapter 3, our goal was to evaluate the practical utility of resources for lexical entailment rules. That is, instead of assessing the correctness of the rules derived from them (i.e., a 'rule-based evaluation'), we evaluate the actual

[1] http://www.nist.gov/tac/2011/RTE/

utility of these resources when used in practical inference scenarios, based on rule applications on actual data instances. That is, an 'instance-based evaluation'.

To that end, we designed an evaluation methodology which operates within an IR setting. Briefly, we make use of short IR queries as hypotheses, e.g. *water pollution*. We modify each hypothesis through the application of entailment rules fetched from the assessed rulebase. This action results with hypotheses such as *lake pollution* or *soil pollution*, following the application of the rules '*lake* $\Rightarrow$ *water*' or '**soil* $\Rightarrow$ *water*', respectively. Then, each modified hypothesis is used as a search query and the retrieved texts are assessed by human annotators. The annotators judge whether texts that entail the modified hypothesis also entail the original one. Whenever this is the case, we consider the rule application successful. Thus, each judged instance represents an application of a single inference operation. Repeating this process quantifies the practical utility of resources for lexical entailment rules. Indeed, the actual figures depend on the setting in use; yet, the comparative results provide some interesting insights with respect to the relative accuracy and coverage of the various resources, as well as for the reasons behind rule application errors.

2.3 Analysis methodology for discourse references

The RTE-5 Search task brought discourse to the spotlight. Up until then, most inference systems and inference-based applications had settled for the simplest means of addressing discourse – through the substitution of nominal coreferents, relying on available tools for coreference resolution.

Based on the understanding presented in Section 1.2.1, that entailment systems share a common goal of obtaining maximal coverage of H by T, we designed an analysis methodology to assess the potential contribution of discourse references (coreference and bridging) to TE-based inference.

The analysis was carried out over a sample of positive examples from the RTE-5 development set, i.e. $T - H$ pairs where T entails H. For each pair, the goal was to identify components in H which cannot be covered directly by T, but rather rely on additional sentences in T's discourse. In Example 2.1 below, none of *AS-28*, *mini* or *underwater* are covered directly by T. We thus search the discourse for references to text expressions in T which can help obtain such coverage. For instance, finding an additional sentence in which *AS-28* is mentioned explicitly and is coreferring with *small submarine* in T will achieve this goal.

(2.1) T : *The Russian military was racing against time early Friday to rescue a small submarine that had become trapped on the seabed with seven crew aboard, the Ria-Novosti news agency reported.*

H : *The AS-28 mini submarine was trapped underwater.*

Indeed, some of the coverage that can be accomplished by means of discourse references resolution can also be obtained through entailment rules, e.g. '*small* $\Leftrightarrow$ *mini*' for covering *mini* in Example 2.1. However, as we could not assume the existence of complete inference knowledge, we conducted the analysis under a no-knowledge assumption. Our results, therefore, provide an "upper-bound" of the potential contribution of discourse references to inference. At the same time, we log those relations that are necessary for correct inference and cannot be resolved without the reliance on discourse, as in the case of anaphoric pronouns. This set of relations constitutes a "lower-bound" of the contribution of discourse references to inference. Among other things, we mark in our analysis the types of discourse references that were identified as relevant for inference (coreference vs. bridging), the parts of speech of participating expressions and the suggested operation that an inference system can perform in order to integrate the reference into the inference process. See Chapter 6 for more details about this analysis methodology and its outcomes.

Chapter 3

Evaluating the Inferential Utility of Lexical-Semantic Resources

Evaluating the Inferential Utility of Lexical-Semantic Resources

Shachar Mirkin, Ido Dagan, Eyal Shnarch
Computer Science Department, Bar-Ilan University
Ramat-Gan 52900, Israel
{mirkins,dagan,shey}@cs.biu.ac.il

Abstract

Lexical-semantic resources are used extensively for applied semantic inference, yet a clear quantitative picture of their current utility and limitations is largely missing. We propose system- and application-independent evaluation and analysis methodologies for resources' performance, and systematically apply them to seven prominent resources. Our findings identify the currently limited recall of available resources, and indicate the potential to improve performance by examining non-standard relation types and by distilling the output of distributional methods. Further, our results stress the need to include auxiliary information regarding the lexical and logical contexts in which a lexical inference is valid, as well as its prior validity likelihood.

1 Introduction

Lexical information plays a major role in semantic inference, as the meaning of one term is often inferred form another. Lexical-semantic resources, which provide the needed knowledge for lexical inference, are commonly utilized by applied inference systems (Giampiccolo et al., 2007) and applications such as Information Retrieval and Question Answering (Shah and Croft, 2004; Pasca and Harabagiu, 2001). Beyond WordNet (Fellbaum, 1998), a wide range of resources has been developed and utilized, including extensions to WordNet (Moldovan and Rus, 2001; Snow et al., 2006) and resources based on automatic distributional similarity methods (Lin, 1998; Pantel and Lin, 2002). Recently, Wikipedia is emerging as a source for extracting semantic relationships (Suchanek et al., 2007; Kazama and Torisawa, 2007).

As of today, only a partial comparative picture is available regarding the actual utility and limitations of available resources for lexical-semantic inference. Works that do provide quantitative information regarding resources utility have focused on few particular resources (Kouylekov and Magnini, 2006; Roth and Sammons, 2007) and evaluated their impact on a specific system. Most often, works which utilized lexical resources do not provide information about their isolated contribution; rather, they only report overall performance for systems in which lexical resources serve as components.

Our paper provides a step towards clarifying this picture. We propose a system- and application-independent evaluation methodology that isolates resources' performance, and systematically apply it to seven prominent lexical-semantic resources. The evaluation and analysis methodology is specified within the Textual Entailment framework, which has become popular in recent years for modeling practical semantic inference in a generic manner (Dagan and Glickman, 2004). To that end, we assume certain definitions that extend the textual entailment paradigm to the lexical level.

The findings of our work provide useful insights and suggested directions for two research communities: developers of applied inference systems and researchers addressing lexical acquisition and resource construction. Beyond the quantitative mapping of resources' performance, our analysis points at issues concerning their effective utilization and major characteristics. Even more importantly, the results highlight current gaps in existing resources and point at directions towards filling them. We show that the coverage of most resources is quite limited, where a substantial part of recall is attributable to semantic relations that are typically not available to inference systems. Notably, distributional acquisition methods

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 558–566,
Athens, Greece, 30 March – 3 April 2009.

are shown to provide many useful relationships which are missing from other resources, but these are embedded amongst many irrelevant ones. Additionally, the results highlight the need to represent and inference over various aspects of contextual information, which affect the applicability of lexical inferences. We suggest that these gaps should be addressed by future research.

2 Sub-sentential Textual Entailment

Textual entailment captures the relation between a *text* t and a textual statement (termed *hypothesis*) h, such that a person reading t would infer that h is most likely correct (Dagan et al., 2005).

The entailment relation has been defined insofar in terms of truth values, assuming that h is a complete sentence (proposition). However, there are major aspects of inference that apply to the sub-sentential level. First, in certain applications, the target hypotheses are often sub-sentential. For example, search queries in IR, which play the hypothesis role from an entailment perspective, typically consist of a single term, like *drug legalization*. Such sub-sentential hypotheses are not regarded naturally in terms of truth values and therefore do not fit well within the scope of the textual entailment definition. Second, many entailment models apply a compositional process, through which they try to infer each sub-part of the hypothesis from some parts of the text (Giampiccolo et al., 2007).

Although inferences over sub-sentential elements are being applied in practice, so far there are no standard definitions for entailment at sub-sentential levels. To that end, and as a prerequisite of our evaluation methodology and our analysis, we first establish two relevant definitions for sub-sentential entailment relations: (a) entailment of a sub-sentential hypothesis by a text, and (b) entailment of one lexical element by another.

2.1 Entailment of Sub-sentential Hypotheses

We first seek a definition that would capture the entailment relationship between a text and a sub-sentential hypothesis. A similar goal was addressed in (Glickman et al., 2006), who defined the notion of *lexical reference* to model the fact that in order to entail a hypothesis, the text has to entail each non-compositional lexical element within it. We suggest that a slight adaptation of their definition is suitable to capture the notion of entailment for any sub-sentential hypotheses, including compositional ones:

Definition 1 *A sub-sentential hypothesis h is entailed by a text t if there is an explicit or implied reference in t to a possible meaning of h.*

For example, the sentence "*crude steel output is likely to fall in 2000*" entails the sub-sentential hypotheses *production*, *steel production* and *steel output decrease*.

Glickman et al., achieving good inter-annotator agreement, empirically found that almost all non-compositional terms in an entailed sentential hypothesis are indeed referenced in the entailing text. This finding suggests that the above definition is consistent with the original definition of textual entailment for sentential hypotheses and can thus model compositional entailment inferences.

We use this definition in our annotation methodology described in Section 3.

2.2 Entailment between Lexical Elements

In the majority of cases, the reference to an "atomic" (non-compositional) lexical element e in h stems from a particular lexical element e' in t, as in the example above where the word *output* implies the meaning of *production*.

To identify this relationship, an entailment system needs a knowledge resource that would specify that the meaning of e' implies the meaning of e, at least in some contexts. We thus suggest the following definition to capture this relationship between e' and e:

Definition 2 *A lexical element e' entails another lexical element e, denoted e'⇒e, if there exist some natural (non-anecdotal) texts containing e' which entail e, such that the reference to the meaning of e can be implied solely from the meaning of e' in the text.*

(Entailment of e by a text follows Definition 1).

We refer to this relation in this paper as *lexical entailment*[1], and call *e' ⇒ e* a *lexical entailment rule*. e' is referred to as the rule's left hand side (*LHS*) and e as its right hand side (*RHS*).

Currently there are no knowledge resources designed specifically for lexical entailment modeling. Hence, the types of relationships they capture do not fully coincide with entailment inference needs. Thus, the definition suggests a specification for the rules that should be provided by

[1] Section 6 discusses other definitions of lexical entailment

a lexical entailment resource, following an operative rationale: a rule *e'* ⇒ *e* should be included in an entailment knowledge resource if it would be needed, as part of a compositional process, to infer the meaning of e from some natural texts. Based on this definition, we perform an analysis of the relationships included in lexical-semantic resources, as described in Section 5.

A rule need not apply in all contexts, as long as it is appropriate for some texts. Two contextual aspects affect rule applicability. First is the "lexical context" specifying the meanings of the text's words. A rules is applicable in a certain context only when the intended sense of its LHS term matches the sense of that term in the text. For example, the application of the rule *lay* ⇒ *produce* is valid only in contexts where the producer is poultry and the products are eggs. This is a well known issue observed, for instance, by Voorhees (1994).

A second contextual factor requiring validation is the "logical context". The logical context determines the monotonicity of the LHS and is induced by logical operators such as negation and (explicit or implicit) quantifiers. For example, the rule *mammal* ⇒ *whale* may not be valid in most cases, but is applicable in universally quantified texts like "*mammals are warm-blooded*". This issue has been rarely addressed in applied inference systems (de Marneffe et al., 2006). The above mentioned rules both comply with Definition 2 and should therefore be included in a lexical entailment resource.

3 Evaluating Entailment Resources

Our evaluation goal is to assess the utility of lexical-semantic resources as sources for entailment rules. An inference system applies a rule by inferring the rule's RHS from texts that match its LHS. Thus, the utility of a resource depends on the performance of its rule applications rather than on the proportion of correct rules it contains. A rule, whether correct or incorrect, has insignificant effect on the resource's utility if it rarely matches texts in real application settings. Additionally, correct rules might produce incorrect applications when applied in inappropriate contexts. Therefore, we use an *instance-based* evaluation methodology, which simulates rule applications by collecting texts that contain rules' LHS and manually assessing the correctness of their applications.

Systems typically handle lexical context either implicitly or explicitly. Implicit context validation occurs when the different terms of a composite hypothesis disambiguate each other. For example, the rule *waterside* ⇒ *bank* is unlikely to be applied when trying to infer the hypothesis *bank loans*, since texts that match *waterside* are unlikely to contain also the meaning of *loan*. Explicit methods, such as word-sense disambiguation or sense matching, validate each rule application according to the broader context in the text. Few systems also address logical context validation by handling quantifiers and negation. As we aim for a system-independent comparison of resources, and explicit approaches are not standardized yet within inference systems, our evaluation uses only implicit context validation.

3.1 Evaluation Methodology

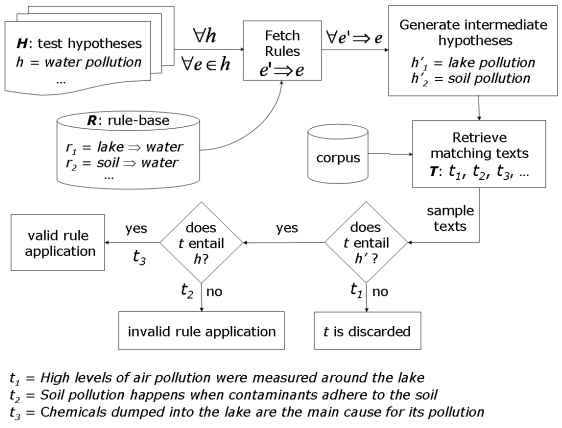

Figure 1: Evaluation methodology flow chart

The input for our evaluation methodology is a lexical-semantic resource R, which contains lexical entailment rules. We evaluate R's utility by testing how useful it is for inferring a sample of test hypotheses H from a corpus. Each hypothesis in H contains more than one lexical element in order to provide implicit context validation for rule applications, e.g. h: *water pollution*.

We next describe the steps of our evaluation methodology, as illustrated in Figure 1. We refer to the examples in the figure when needed:

1) Fetch rules: For each $h \in H$ and each lexical element $e \in h$ (e.g. *water*), we fetch all rules *e'* ⇒ *e* in R that might be applied to entail e (e.g. *lake* ⇒ *water*).

2) Generate intermediate hypotheses h': For each rule r: *e'* ⇒ *e*, we generate an intermediate hypothesis h' by replacing e in h with e' (e.g.

h'_1: *lake pollution*). From a text t entailing h', h can be further entailed by the single application of r. We thus simulate the process by which an entailment system would infer h from t using r.

3) Retrieve matching texts: For each h' we retrieve from a corpus all texts that contain the lemmatized words of h' (not necessarily as a single phrase). These texts may entail h'. We discard texts that also match h since entailing h from them might not require the application of any rule from the evaluated resource. In our example, the retrieved texts contain *lake* and *pollution* but do not contain *water*.

4) Annotation: A sample of the retrieved texts is presented to human annotators. The annotators are asked to answer the following two questions for each text, simulating the typical inference process of an entailment system:

a) Does t entail h'? If t does not entail h' then the text would not provide a useful example for the application of r. For instance, t_1 (in Figure 1) does not entail h'_1 and thus we cannot deduce h from it by applying the rule r. Such texts are discarded from further evaluation.

b) Does t entail h? If t is annotated as entailing h', an entailment system would then infer h from h' by applying r. If h is not entailed from t even though h' is, the rule application is considered invalid. For instance, t_2 does not entail h even though it entails h'_2. Indeed, the application of r_2: **soil ⇒ water* [2], from which h'_2 was constructed, yields incorrect inference. If the answer is 'yes', as in the case of t_3, the application of r for t is considered valid.

The above process yields a sample of annotated rule applications for each test hypothesis, from which we can measure resources performance, as described in Section 5.

4 Experimental Setting

4.1 Dataset and Annotation

Current available state-of-the-art lexical-semantic resources mainly deal with nouns. Therefore, we used nominal hypotheses for our experiment[3].

We chose TREC 1-8 (excluding 4) as our test corpus and randomly sampled 25 ad-hoc queries of two-word compounds as our hypotheses. We did not use longer hypotheses to ensure that enough texts containing the intermediate hypotheses are found in the corpus. For annotation simplicity, we retrieved single sentences as our texts.

For each rule applied for an hypothesis h, we sampled 10 sentences from the sentences retrieved for that rule. As a baseline, we also sampled 10 sentences for each original hypothesis h in which both words of h are found. In total, 1550 unique sentences were sampled and annotated by two annotators.

To assess the validity of our evaluation methodology, the annotators first judged a sample of 220 sentences. The Kappa scores for inter-annotator agreement were 0.74 and 0.64 for judging h' and h, respectively. These figures correspond to substantial agreement (Landis and Koch, 1997) and are comparable with related semantic annotations (Szpektor et al., 2007; Bhagat et al., 2007).

4.2 Lexical-Semantic Resources

We evaluated the following resources:

WordNet (*WN*^{d}): There is no clear agreement regarding which set of WordNet relations is useful for entailment inference. We therefore took a conservative approach using only synonymy and hyponymy rules, which typically comply with the lexical entailment relation and are commonly used by textual entailment systems, e.g. (Herrera et al., 2005; Bos and Markert, 2006). Given a term e, we created a rule *e'* ⇒ *e* for each e' amongst the synonyms or direct hyponyms for all senses of e in WordNet 3.0.

Snow (*Snow*^{30k}): Snow et al. (2006) presented a probabilistic model for taxonomy induction which considers as features paths in parse trees between related taxonomy nodes. They show that the best performing taxonomy was the one adding 30,000 hyponyms to WordNet. We created an entailment rule for each new hyponym added to WordNet by their algorithm[4].

LCC's extended WordNet (*XWN**): In (Moldovan and Rus, 2001) WordNet glosses were transformed into logical form axioms. From this representation we created a rule *e'* ⇒ *e* for each e' in the gloss which was tagged as referring to the same entity as e.

CBC: A knowledgebase of labeled clusters generated by the statistical clustering and labeling algorithms in (Pantel and Lin, 2002; Pantel and

[2] The asterisk marks an incorrect rule.

[3] We suggest that the definitions and methodologies can be applied for other parts of speech as well.

[4] Available at http://ai.stanford.edu/~rion/swn

Ravichandran, 2004)[5]. Given a cluster label e, an entailment rule *e'* ⇒ *e* is created for each member e' of the cluster.

Lin Dependency Similarity (*Lin-dep*): A distributional word similarity resource based on syntactic-dependency features (Lin, 1998). Given a term e and its list of similar terms, we construct for each e' in the list the rule *e'* ⇒ *e*. This resource was previously used in textual entailment engines, e.g. (Roth and Sammons, 2007).

Lin Proximity Similarity (*Lin-prox*): A knowledgebase of terms with their cooccurrence-based distributionally similar terms. Rules are created from this resource as from the previous one[6].

Wikipedia first sentence (*WikiFS*): Kazama and Torisawa (2007) used Wikipedia as an external knowledge to improve Named Entity Recognition. Using the first step of their algorithm, we extracted from the first sentence of each page a noun that appears in a *is-a* pattern referring to the title. For each such pair we constructed a rule *title* ⇒ *noun* (e.g. *Michelle Pfeiffer* ⇒ *actress*).

The above resources represent various methods for detecting semantic relatedness between words: Manually and semi-automatically constructed (*WN*^{d} and *XWN**, respectively), automatically constructed based on a lexical-syntactic pattern (*WikiFS*), distributional methods (*Lin-dep* and *Lin-prox*) and combinations of pattern-based and distributional methods (*CBC* and *Snow*^{30k}).

5 Results and Analysis

The results and analysis described in this section reveal new aspects concerning the utility of resources for lexical entailment, and experimentally quantify several intuitively-accepted notions regarding these resources and the lexical entailment relation. Overall, our findings highlight where efforts in developing future resources and inference systems should be invested.

5.1 Resources Performance

Each resource was evaluated using two measures - *Precision* and *Recall-share*, macro averaged over all hypotheses. The results achieved for each resource are summarized in Table 1.

[5]Kindly provided to us by Patrick Pantel.

[6]Lin's resources were downloaded from: http://www.cs.ualberta.ca/~lindek/demos.htm

Resource	Precision (%)	Recall-share (%)
Snow^{30k}	56	8
WN^d	55	24
XWN*	51	9
WikiFS	45	7
CBC	33	9
Lin-dep	28	45
Lin-prox	24	36

Table 1: Lexical resources performance

5.1.1 Precision

The *Precision* of a resource R is the percentage of valid rule applications for the resource. It is estimated by the percentage of texts entailing h from those that entail h': $\frac{count_R(entailing\ h{=}yes)}{count_R(entailing\ h'{=}yes)}$.

Not surprisingly, resources such as *WN*^{d}, *XWN** or *WikiFS* achieved relatively high precision scores, due to their accurate construction methods. In contrast, Lin's distributional resources are not designed to include lexical entailment relationships. They provide pairs of contextually similar words, of which many have non-entailing relationships, such as co-hyponyms[7] (e.g. **doctor* ⇒ *journalist*) or topically-related words, such as **radiotherapy* ⇒ *outpatient*. Hence their relatively low precision.

One visible outcome is the large gap between the perceived high accuracy of resources constructed by accurate methods, most notably *WN*^{d}, and their performance in practice. This finding emphasizes the need for instance-based evaluations, which capture the "real" contribution of a resource. To better understand the reasons for this gap we further assessed the three factors that contribute to incorrect applications: incorrect rules, lexical context and logical context (see Section 2.2). This analysis is presented in Table 2.

From Table 2 we see that the gap for accurate resources is mainly caused by applications of correct rules in inappropriate contexts. More interestingly, the information in the table allows us to asses the lexical "context-sensitivity" of resources. When considering only the COR-LEX rules to recalculate resources precision, we find that *Lin-dep* achieves precision of 71% ($\frac{15\%}{15\%+6\%}$), while *WN*^{d} yields only 56% ($\frac{55\%}{55\%+44\%}$). This result indicates that correct *Lin-dep* rules are less sensitive to lexical context, meaning that their prior likelihoods to

[7]a.k.a. *sister terms* or *coordinate terms*

(%)	Invalid Rule Applications				Valid Rule Applications			
	INCOR	COR-LOG	COR-LEX	Total	INCOR	COR-LOG	COR-LEX	Total (P)
WN^d	1	0	44	45	0	0	55	55
WikiFS	13	0	42	55	3	0	42	45
XWN*	19	0	30	49	0	0	51	51
Snow^{30k}	23	0	21	44	0	0	56	56
CBC	51	12	4	67	14	0	19	33
Lin-prox	59	4	13	76	8	3	13	24
Lin-dep	61	5	6	72	9	4	15	28

Table 2: The distribution of invalid and valid rule applications by rule types: incorrect rules (INCOR), correct rules requiring "logical context" validation (COR-LOG), and correct rules requiring "lexical context" matching (COR-LEX). The numbers of each resource's valid applications add up to the resource's precision.

be correct are higher. This is explained by the fact that *Lin-dep*'s rules are calculated across multiple contexts and therefore capture the more frequent usages of words. WordNet, on the other hand, includes many anecdotal rules whose application is rare, and thus is very sensitive to context. Similarly, *WikiFS* turns out to be very context-sensitive. This resource contains many rules for polysemous proper nouns that are scarce in their proper noun sense, e.g. *Captive* $\Rightarrow$ *computer game*. *Snow^{30k}*, when applied with the same calculation, reaches 73%, which explains how it achieved a comparable result to *WN^d*, even though it contains many incorrect rules in comparison to *WN^d*.

5.1.2 Recall

Absolute recall cannot be measured since the total number of texts in the corpus that entail each hypothesis is unknown. Instead, we measure *recall-share*, the contribution of each resource to recall relative to matching only the words of the original hypothesis without any rules. We denote by $yield(h)$ the number of texts that match h directly and are annotated as entailing h. This figure is estimated by the number of sampled texts annotated as entailing h multiplied by the sampling proportion. In the same fashion, for each resource R, we estimate the number of texts entailing h obtained through entailment rules of the resource R, denoted $yield_R(h)$. Recall-share of R for h is the proportion of the yield obtained by the resource's rules relative to the overall yield with and without the rules from R: $\frac{yield_R(h)}{yield(h)+yield_R(h)}$.

From Table 1 we see that along with their relatively low precision, Lin's resources' recall greatly surpasses that of any other resource, including WordNet[8]. The rest of the resources are even inferior to *WN^d* in that respect, indicating their limited utility for inference systems.

As expected, synonyms and hyponyms in WordNet contributed a noticeable portion to recall in all resources. Additional correct rules correspond to hyponyms and synonyms missing from WordNet, many of them proper names and some slang expressions. These rules were mainly provided by *WikiFS* and *Snow^{30k}*, significantly supplementing WordNet, whose *HasInstance* relation is quite partial. However, there are other interesting types of entailment relations contributing to recall. These are discussed in Sections 5.2 and 5.3. Examples for various rule types are found in Table 3.

5.1.3 Valid Applications of Incorrect Rules

We observed that many entailing sentences were retrieved by inherently incorrect rules in the distributional resources. Analysis of these rules reveals they were matched in entailing texts when the LHS has noticeable statistical correlation with another term in the text that does entail the RHS. For example, for the hypothesis *wildlife extinction*, the rule **species* $\Rightarrow$ *extinction* yielded valid applications in contexts about *threatened* or *endangered species*. Has the resource included a rule between the entailing term in the text and the RHS, the entailing text would have been matched without needing the incorrect rule.

These correlations accounted for nearly a third of Lin resources' recall. Nonetheless, in principle, we suggest that such rules, which do not conform with Definition 2, should not be included in a lexical entailment resource, since they also cause invalid rule applications, while the entailing texts they retrieve will hopefully be matched by addi-

[8] A preliminary experiment we conducted showed that recall does not dramatically improve when using the entire hyponymy subtree from WordNet.

Type	Correct Rules	
HYPO	*Shevardnadze* ⇒ *official*	*Snow*^{30k}
ANT	*efficacy* ⇒ *ineffectiveness*	*Lin-dep*
HOLO	*government* ⇒ *official*	*Lin-prox*
HYPER	*arms* ⇒ *gun*	*Lin-prox*
-	*childbirth* ⇒ *motherhood*	*Lin-dep*
-	*mortgage* ⇒ *bank*	*Lin-prox*
-	*Captive* ⇒ *computer*	*WikiFS*
-	*negligence* ⇒ *failure*	*CBC*
-	*beatification* ⇒ *pope*	*XWN**

Type	Incorrect Rules	
CO-HYP	*alcohol* ⇒ *cigarette*	*CBC*
-	*radiotherapy* ⇒ *outpatient*	*Lin-dep*
-	*teen-ager* ⇒ *gun*	*Snow*^{30k}
-	*basic* ⇒ *paper*	*WikiFS*
-	*species* ⇒ *extinction*	*Lin-prox*

Table 3: Examples of lexical resources rules by types. HYPO: hyponymy, HYPER: hypernymy (class entailment of its members), HOLO: holonymy, ANT: antonymy, CO-HYP: co-hyponymy. The non-categorized relations do not correspond to any WordNet relation.

tional correct rules in a more comprehensive resource.

5.2 Non-standard Entailment Relations

An important finding of our analysis is that some less standard entailment relationships have a considerable impact on recall (see Table 3). These rules, which comply with Definition 2 but do not conform to any WordNet relation type, were mainly contributed by Lin's distributional resources and to a smaller degree are also included in *XWN**. In *Lin-dep*, for example, they accounted for approximately a third of the recall.

Among the finer grained relations we identified in this set are topical entailment (e.g. *IBM* as the company entailing the topic *computers*), consequential relationships (*pregnancy* ⇒ *motherhood*) and an entailment of inherent arguments by a predicate, or of essential participants by a scenario description, e.g. *beatification* ⇒ *pope*. A comprehensive typology of these relationships requires further investigation, as well as the identification and development of additional resources from which they can be extracted.

As opposed to hyponymy and synonymy rules, these rules are typically non-substitutable, i.e. the RHS of the rule is unlikely to have the exact same role in the text as the LHS. Many inference systems perform rule-based transformations, substituting the LHS by the RHS. This finding suggests that different methods may be required to utilize such rules for inference.

5.3 Logical Context

WordNet relations other than synonyms and hyponyms, including antonyms, holonyms and hypernyms (see Table 3), contributed a noticeable share of valid rule applications for some resources. Following common practice, these relations are missing by construction from the other resources.

As shown in Table 2 (COR-LOG columns), such relations accounted for a seventh of *Lin-dep*'s valid rule applications, as much as was the contribution of hyponyms and synonyms to this resource's recall. Yet, using these rules resulted with more erroneous applications than correct ones. As discussed in Section 2.2, the rules induced by these relations do conform with our lexical entailment definition. However, a valid application of these rules requires certain logical conditions to occur, which is not the common case. We thus suggest that such rules are included in lexical entailment resources, as long as they are marked properly by their types, allowing inference systems to utilize them only when appropriate mechanisms for handling logical context are in place.

5.4 Rules Priors

In Section 5.1.1 we observed that some resources are highly sensitive to context. Hence, when considering the validity of a rule's application, two factors should be regarded: the actual context in which the rule is to be applied, as well as the rule's prior likelihood to be valid in an arbitrary context. Somewhat indicative, yet mostly indirect, information about rules' priors is contained in some resources. This includes sense ranks in WordNet, SemCor statistics (Miller et al., 1993), and similarity scores and rankings in Lin's resources. Inference systems often incorporated this information, typically as top-*k* or threshold-based filters (Pantel and Lin, 2003; Roth and Sammons, 2007). By empirically assessing the effect of several such filters in our setting, we found that this type of data is indeed informative in the sense that precision increases as the threshold rises. Yet, no specific filters were found to improve results in terms of F1 score (where recall is measured relatively to the yield of the unfiltered resource) due to a significant drop in relative recall. For example, *Lin-*

prox loses more than 40% of its recall when only the top-50 rules for each hypothesis are exploited, and using only the first sense of WN^d costs the resource over 60% of its recall. We thus suggest a better strategy might be to combine the prior information with context matching scores in order to obtain overall likelihood scores for rule applications, as in (Szpektor et al., 2008). Furthermore, resources should include explicit information regarding the prior likelihoods of of their rules.

5.5 Operative Conclusions

Our findings highlight the currently limited recall of available resources for lexical inference. The higher recall of Lin's resources indicates that many more entailment relationships can be acquired, particularly when considering distributional evidence. Yet, available distributional acquisition methods are not geared for lexical entailment. This suggests the need to develop acquisition methods for dedicated and more extensive knowledge resources that would subsume the correct rules found by current distributional methods. Furthermore, substantially better recall may be obtained by acquiring non-standard lexical entailment relationships, as discussed in Section 5.2, for which a comprehensive typology is still needed. At the same time, transformation-based inference systems would need to handle these kinds of rules, which are usually non-substitutable. Our results also quantify and stress earlier findings regarding the severe degradation in precision when rules are applied in inappropriate contexts. This highlights the need for resources to provide explicit information about the suitable lexical and logical contexts in which an entailment rule is applicable. In parallel, methods should be developed to utilize such contextual information within inference systems. Additional auxiliary information needed in lexical resources is the prior likelihood for a given rule to be correct in an arbitrary context.

6 Related Work

Several prior works defined lexical entailment. WordNet's lexical entailment is a relationship between verbs only, defined for propositions (Fellbaum, 1998). Geffet and Dagan (2004) defined *substitutable lexical entailment* as a relation between substitutable terms. We find this definition too restrictive as non-substitutable rules may also be useful for entailment inference. Examples are *breastfeeding* $\Rightarrow$ *baby* and *hospital* $\Rightarrow$ *medical*. Hence, Definition 2 is more broadly applicable for defining the desired contents of lexical entailment resources. We empirically observed that the rules satisfying their definition are a proper subset of the rules covered by our definition. Dagan and Glickman (2004) referred to entailment at the subsentential level by assigning truth values to subpropositional text fragments through their existential meaning. We find this criterion too permissive. For instance, the existence of *country* implies the existence of its *flag*. Yet, the meaning of *flag* is typically not implied by *country*.

Previous works assessing rule application via human annotation include (Pantel et al., 2007; Szpektor et al., 2007), which evaluate acquisition methods for lexical-syntactic rules. They posed an additional question to the annotators asking them to filter out invalid contexts. In our methodology implicit context matching for the full hypothesis was applied instead. Other related instance-based evaluations (Giuliano and Gliozzo, 2007; Connor and Roth, 2007) performed lexical substitutions, but did not handle the non-substitutable cases.

7 Conclusions

This paper provides several methodological and empirical contributions. We presented a novel evaluation methodology for the utility of lexical-semantic resources for semantic inference. To that end we proposed definitions for entailment at subsentential levels, addressing a gap in the textual entailment framework. Our evaluation and analysis provide a first quantitative comparative assessment of the isolated utility of a range of prominent potential resources for entailment rules. We have shown various factors affecting rule applicability and resources performance, while providing operative suggestions to address them in future inference systems and resources.

Acknowledgments

The authors would like to thank Naomi Frankel and Iddo Greental for their excellent annotation work, as well as Roy Bar-Haim and Idan Szpektor for helpful discussion and advice. This work was partially supported by the Negev Consortium of the Israeli Ministry of Industry, Trade and Labor, the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886 and the Israel Science Foundation grant 1095/05.

References

Rahul Bhagat, Patrick Pantel, and Eduard Hovy. 2007. LEDIR: An unsupervised algorithm for learning directionality of inference rules. In *Proceedings of EMNLP-CoNLL*.

J. Bos and K. Markert. 2006. When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of the Second PASCAL RTE Challenge*.

Michael Connor and Dan Roth. 2007. Context sensitive paraphrasing with a global unsupervised classifier. In *Proceedings of ECML*.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *PASCAL Workshop on Learning Methods for Text Understanding and Mining*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In Joaquin Quinonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *MLCW*, Lecture Notes in Computer Science.

Marie-Catherine de Marneffe, Bill MacCartney, Trond Grenager, Daniel Cer, Anna Rafferty, and Christopher D. Manning. 2006. Learning to distinguish valid textual entailments. In *Proceedings of the Second PASCAL RTE Challenge*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Maayan Geffet and Ido Dagan. 2004. Feature vector quality and distributional similarity. In *Proceedings of COLING*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of ACL-WTEP Workshop*.

Claudio Giuliano and Alfio Gliozzo. 2007. Instance based lexical entailment for ontology population. In *Proceedings of EMNLP-CoNLL*.

Oren Glickman, Eyal Shnarch, and Ido Dagan. 2006. Lexical reference: a semantic matching subtask. In *Proceedings of EMNLP*.

Jesús Herrera, Anselmo Peñas, and Felisa Verdejo. 2005. Textual entailment recognition based on dependency analysis and wordnet. In *Proceedings of the First PASCAL RTE Challenge*.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of EMNLP-CoNLL*.

Milen Kouylekov and Bernardo Magnini. 2006. Building a large-scale repository of textual entailment rules. In *Proceedings of LREC*.

J. R. Landis and G. G. Koch. 1997. The measurements of observer agreement for categorical data. In *Biometrics*, pages 33:159–174.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of HLT*.

Dan Moldovan and Vasile Rus. 2001. Logic form transformation of wordnet and its applicability to question answering. In *Proceedings of ACL*.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD*.

Patrick Pantel and Dekang Lin. 2003. Automatically discovering word senses. In *Proceedings of NAACL*.

Patrick Pantel and Deepak Ravichandran. 2004. Automatically labeling semantic classes. In *Proceedings of HLT-NAACL*.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Proceedings of HLT*.

Marius Pasca and Sanda M. Harabagiu. 2001. The informative role of wordnet in open-domain question answering. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*.

Dan Roth and Mark Sammons. 2007. Semantic and logical inference model for textual entailment. In *Proceedings of ACL-WTEP Workshop*.

Chirag Shah and Bruce W. Croft. 2004. Evaluating high accuracy retrieval techniques. In *Proceedings of SIGIR*.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING-ACL*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge - unifying wordnet and wikipedia. In *Proceedings of WWW*.

Idan Szpektor, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of ACL*.

Idan Szpektor, Ido Dagan, Roy Bar-Haim, and Jacob Goldberger. 2008. Contextual preferences. In *Proceedings of ACL*.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of SIGIR*.

Chapter 4

Source-Language Entailment Modeling for Translating Unknown Terms

Source-Language Entailment Modeling for Translating Unknown Terms

Shachar Mirkin[§], **Lucia Specia**[†], **Nicola Cancedda**[†], **Ido Dagan**[§], **Marc Dymetman**[†], **Idan Szpektor**[§]
§ Computer Science Department, Bar-Ilan University
† Xerox Research Centre Europe
{mirkins,dagan,szpekti}@cs.biu.ac.il
{lucia.specia,nicola.cancedda,marc.dymetman}@xrce.xerox.com

Abstract

This paper addresses the task of handling unknown terms in SMT. We propose using source-language monolingual models and resources to paraphrase the source text prior to translation. We further present a conceptual extension to prior work by allowing translations of *entailed* texts rather than paraphrases only. A method for performing this process efficiently is presented and applied to some 2500 sentences with unknown terms. Our experiments show that the proposed approach substantially increases the number of properly translated texts.

1 Introduction

Machine Translation systems frequently encounter terms they are not able to translate due to some missing knowledge. For instance, a Statistical Machine Translation (SMT) system translating the sentence "*Cisco filed a lawsuit against Apple for patent violation*" may lack words like *filed* and *lawsuit* in its phrase table. The problem is especially severe for languages for which parallel corpora are scarce, or in the common scenario when the SMT system is used to translate texts of a domain different from the one it was trained on.

A previously suggested solution (Callison-Burch et al., 2006) is to learn paraphrases of source terms from multilingual (parallel) corpora, and expand the phrase table with these paraphrases[1]. Such solutions could potentially yield a paraphrased sentence like "*Cisco sued Apple for patent violation*", although their dependence on bilingual resources limits their utility.

In this paper we propose an approach that consists in directly replacing unknown source terms, using source-language resources and models in order to achieve two goals.

The first goal is coverage increase. The availability of bilingual corpora, from which paraphrases can be learnt, is in many cases limited. On the other hand, monolingual resources and methods for extracting paraphrases from monolingual corpora are more readily available. These include manually constructed resources, such as WordNet (Fellbaum, 1998), and automatic methods for paraphrases acquisition, such as DIRT (Lin and Pantel, 2001). However, such resources have not been applied yet to the problem of substituting unknown terms in SMT. We suggest that by using such monolingual resources we could provide paraphrases for a larger number of texts with unknown terms, thus increasing the overall coverage of the SMT system, i.e. the number of texts it properly translates.

Even with larger paraphrase resources, we may encounter texts in which not all unknown terms are successfully handled through paraphrasing, which often results in poor translations (see Section 2.1). To further increase coverage, we therefore propose to generate and translate texts that convey a somewhat more general meaning than the original source text. For example, using such approach, the following text could be generated: "*Cisco accused Apple of patent violation*". Although less informative than the original, a translation for such texts may be useful. Such non-symmetric relationships (as between *filed a lawsuit* and *accused*) are difficult to learn from parallel corpora and therefore monolingual resources are more appropriate for this purpose.

The second goal we wish to accomplish by employing source-language resources is to rank the alternative generated texts. This goal can be achieved by using context-models on the source language prior to translation. This has two advantages. First, the ranking allows us to prune some

[1] As common in the literature, we use the term *paraphrases* to refer to texts of equivalent meaning, of any length from single words (synonyms) up to complete sentences.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 791–799,
Suntec, Singapore, 2-7 August 2009.

candidates before supplying them to the translation engine, thus improving translation efficiency. Second, the ranking may be combined with target language information in order to choose the best translation, thus improving translation quality.

We position the problem of generating alternative texts for translation within the Textual Entailment (TE) framework (Giampiccolo et al., 2007). TE provides a generic way for handling language variability, identifying when the meaning of one text is *entailed* by the other (i.e. the meaning of the entailed text can be inferred from the meaning of the entailing one). When the meanings of two texts are equivalent (paraphrase), entailment is mutual. Typically, a more general version of a certain text is entailed by it. Hence, through TE we can formalize the generation of both equivalent and more general texts for the source text. When possible, a paraphrase is used. Otherwise, an alternative text whose meaning is entailed by the original source is generated and translated.

We assess our approach by applying an SMT system to a text domain that is different from the one used to train the system. We use WordNet as a source language resource for entailment relationships and several common statistical context-models for selecting the best generated texts to be sent to translation. We show that the use of source language resources, and in particular the extension to non-symmetric textual entailment relationships, is useful for substantially increasing the amount of texts that are properly translated. This increase is observed relative to both using paraphrases produced by the same resource (WordNet) and using paraphrases produced from multilingual parallel corpora. We demonstrate that by using simple context-models on the source, efficiency can be improved, while translation quality is maintained. We believe that with the use of more sophisticated context-models further quality improvement can be achieved.

2 Background

2.1 Unknown Terms

A very common problem faced by machine translation systems is the need to translate terms (words or multi-word expressions) that are not found in the system's lexicon or phrase table. The reasons for such unknown terms in SMT systems include scarcity of training material and the application of the system to text domains that differ from the ones used for training.

In SMT, when unknown terms are found in the source text, the systems usually omit or copy them literally into the target. Though copying the source words can be of some help to the reader if the unknown word has a cognate in the target language, this will not happen in the most general scenario where, for instance, languages use different scripts. In addition, the presence of a single unknown term often affects the translation of wider portions of text, inducing errors in both lexical selection and ordering. This phenomenon is demonstrated in the following sentences, where the translation of the English sentence (1) is acceptable only when the unknown word (in bold) is replaced with a translatable paraphrase (3):

1. "*..., despite bearing the heavy burden of the unemployed 10% or more of the* ***labor*** *force.*"
2. "*..., malgré la lourde charge de compte le 10% ou plus de chômeurs* ***labor*** *la force .*"
3. "*..., malgré la lourde charge des chômeurs de 10% ou plus de la force du* ***travail****.*"

Several approaches have been proposed to deal with unknown terms in SMT systems, rather than omitting or copying the terms. For example, (Eck et al., 2008) replace the unknown terms in the source text by their definition in a monolingual dictionary, which can be useful for gisting. To translate across languages with different alphabets approaches such as (Knight and Graehl, 1997; Habash, 2008) use transliteration techniques to tackle proper nouns and technical terms. For translation from highly inflected languages, certain approaches rely on some form of lexical approximation or morphological analysis (Koehn and Knight, 2003; Yang and Kirchhoff, 2006; Langlais and Patry, 2007; Arora et al., 2008). Although these strategies yield gain in coverage and translation quality, they only account for unknown terms that should be transliterated or are variations of known ones.

2.2 Paraphrasing in MT

A recent strategy to broadly deal with the problem of unknown terms is to paraphrase the source text with terms whose translation is known to the system, using paraphrases learnt from multilingual corpora, typically involving at least one "pivot" language different from the target language of immediate interest (Callison-Burch et

al., 2006; Cohn and Lapata, 2007; Zhao et al., 2008; Callison-Burch, 2008; Guzmán and Garrido, 2008). The procedure to extract paraphrases in these approaches is similar to standard phrase extraction in SMT systems, and therefore a large amount of additional parallel corpus is required. Moreover, as discussed in Section 5, when unknown texts are not from the same domain as the SMT training corpus, it is likely that paraphrases found through such methods will yield misleading translations.

Bond et al. (2008) use grammars to paraphrase the whole source sentence, covering aspects like word order and minor lexical variations (tenses etc.), but not content words. The paraphrases are added to the source side of the corpus and the corresponding target sentences are duplicated. This, however, may yield distorted probability estimates in the phrase table, since these were not computed from parallel data.

The main use of monolingual paraphrases in MT to date has been for evaluation. For example, (Kauchak and Barzilay, 2006) paraphrase references to make them closer to the system translation in order to obtain more reliable results when using automatic evaluation metrics like BLEU (Papineni et al., 2002).

2.3 Textual Entailment and Entailment Rules

Textual Entailment (TE) has recently become a prominent paradigm for modeling semantic inference, capturing the needs of a broad range of text understanding applications (Giampiccolo et al., 2007). Yet, its application to SMT has been so far limited to MT evaluation (Pado et al., 2009).

TE defines a directional relation between two texts, where the meaning of the entailed text (*hypothesis*, *h*) can be inferred from the meaning of the entailing *text, t*. Under this paradigm, paraphrases are a special case of the entailment relation, when the relation is symmetric (the texts entail each other). Otherwise, we say that one text *directionally entails* the other.

A common practice for proving (or generating) *h* from *t* is to apply *entailment rules* to *t*. An entailment rule, denoted *LHS* $\Rightarrow$ *RHS*, specifies an entailment relation between two text fragments (the Left- and Right- Hand Sides), possibly with variables (e.g. *build* X *in* $Y \Rightarrow X$ *is completed in* Y). A paraphrasing rule is denoted with $\Leftrightarrow$. When a rule is applied to a text, a new text is inferred, where the matched LHS is replaced with the RHS. For example, the rule *skyscraper* $\Rightarrow$ *building* is applied to "*The world's tallest skyscraper was completed in Taiwan*" to infer "*The world's tallest building was completed in Taiwan*". In this work, we employ lexical entailment rules, i.e. rules without variables. Various resources for lexical rules are available, and the prominent one is WordNet (Fellbaum, 1998), which has been used in virtually all TE systems (Giampiccolo et al., 2007).

Typically, a rule application is valid only under specific contexts. For example, *mouse* $\Rightarrow$ *rodent* should not be applied to "*Use the mouse to mark your answers*". Context-models can be exploited to validate the application of a rule to a text. In such models, an explicit Word Sense Disambiguation (WSD) is not necessarily required; rather, an implicit sense-match is sought after (Dagan et al., 2006). Within the scope of our paper, rule application is handled similarly to Lexical Substitution (McCarthy and Navigli, 2007), considering the contextual relationship between the text and the rule. However, in general, entailment rule application addresses other aspects of context matching as well (Szpektor et al., 2008).

3 Textual Entailment for Statistical Machine Translation

Previous solutions for handling unknown terms in a source text s augment the SMT system's phrase table based on multilingual corpora. This allows indirectly paraphrasing s, when the SMT system chooses to use a paraphrase included in the table and produces a translation with the corresponding target phrase for the unknown term.

We propose using monolingual paraphrasing methods and resources for this task to obtain a more extensive set of rules for paraphrasing the source. These rules are then applied to s directly to produce alternative versions of the source text prior to the translation step. Moreover, further coverage increase can be achieved by employing directional entailment rules, when paraphrasing is not possible, to generate more general texts for translation.

Our approach, based on the textual entailment framework, considers the newly generated texts as entailed from the original one. Monolingual semantic resources such as WordNet can provide entailment rules required for both these symmetric and asymmetric entailment relations.

Input: A text t with one or more unknown terms;
a monolingual resource of entailment rules;
k - maximal number of source alternatives to produce

Output: A translation of either (in order of preference):
a paraphrase of t OR a text entailed by t OR t itself

1. For each unknown term - fetch entailment rules:
   (a) Fetch rules for paraphrasing; disregard rules whose RHS is not in the phrase table
   (b) If the set of rules is empty: fetch directional entailment rules; disregard rules whose RHS is not in the phrase table
2. Apply a context-model to compute a score for each rule application
3. Compute total source score for each entailed text as a combination of individual rule scores
4. Generate and translate the top-*k* entailed texts
5. If $k > 1$
   (a) Apply target model to score the translation
   (b) Compute final source-target score
6. Pick highest scoring translation

Figure 1: Scheme for handling unknown terms by using monolingual resources through textual entailment

Through the process of applying entailment rules to the source text, multiple alternatives of entailed texts are generated. To rank the candidate texts we employ monolingual context-models to provide scores for rule applications over the source sentence. This can be used to (a) directly select the text with the highest score, which can then be translated, or (b) to select a subset of top candidates to be translated, which will then be ranked using the target language information as well. This pruning reduces the load of the SMT system, and allows for potential improvements in translation quality by considering both source- and target-language information.

The general scheme through which we achieve these goals, which can be implemented using different context-models and scoring techniques, is detailed in Figure 1. Details of our concrete implementation are given in Section 4.

Preliminary analysis confirmed (as expected) that readers prefer translations of paraphrases, when available, over translations of directional entailments. This consideration is therefore taken into account in the proposed method.

The input is a text unit to be translated, such as a sentence or paragraph, with one or more unknown terms. For each unknown term we first fetch a list of candidate rules for paraphrasing (e.g. synonyms), where the unknown term is the LHS. For example, if our unknown term is *dodge*, a possible candidate might be *dodge* $\Leftrightarrow$ *circumvent*. We inflect the RHS to keep the original morphological information of the unknown term and filter out rules where the inflected RHS does not appear in the phrase table (step 1a in Figure 1).

When no applicable rules for paraphrasing are available (1b), we fetch directional entailment rules (e.g. hypernymy rules such as *dodge* $\Rightarrow$ *avoid*), and filter them in the same way as for paraphrasing rules. To each set of rules for a given unknown term we add the "identity-rule", to allow leaving the unknown term unchanged, the correct choice in cases of proper names, for example.

Next, we apply a context-model to compute an applicability score of each rule to the source text (step 2). An entailed text's total score is the combination (e.g. product, see Section 4) of the scores of the rules used to produce it (3). A set of the top-*k* entailed texts is then generated and sent for translation (4).

If more than one alternative is produced by the source model (and $k > 1$), a target model is applied on the selected set of translated texts (5a). The combined source-target model score is a combination of the scores of the source and target models (5b). The final translation is selected to be the one that yields the highest combined source-target score (6). Note that setting $k = 1$ is equivalent to using the source-language model alone.

Our algorithm validates the application of the entailment rules at two stages – before and after translation, through context-models applied at each end. As the experiments will show in Section 4, a large number of possible combinations of entailment rules is a common scenario, and therefore using the source context models to reduce this number plays an important role.

4 Experimental Setting

To assess our approach, we conducted a series of experiments; in each experiment we applied the scheme described in 3, changing only the models being used for scoring the generated and translated texts. The setting of these experiments is described in what follows.

SMT data To produce sentences for our experiments, we use Matrax (Simard et al., 2005), a standard phrase-based SMT system, with the exception that it allows gaps in phrases. We use approximately 1M sentence pairs from the English-French

Europarl corpus for training, and then translate a test set of 5,859 English sentences from the News corpus into French. Both resources are taken from the shared translation task in WMT-2008 (Callison-Burch et al., 2008). Hence, we compare our method in a setting where the training and test data are from different domains, a common scenario in the practical use of MT systems.

Of the 5,859 translated sentences, 2,494 contain unknown terms (considering only sequences with alphabetic symbols), summing up to 4,255 occurrences of unknown terms. 39% of the 2,494 sentences contain more than a single unknown term.

Entailment resource We use WordNet 3.0 as a resource for entailment rules. Paraphrases are generated using synonyms. Directionally entailed texts are created using hypernyms, which typically conform with entailment. We do not rely on sense information in WordNet. Hence, any other semantic resource for entailment rules can be utilized.

Each sentence is tagged using the OpenNLP POS tagger[2]. Entailment rules are applied for unknown terms tagged as nouns, verbs, adjectives and adverbs. The use of relations from WordNet results in 1,071 sentences with applicable rules (with phrase table entries) for the unknown terms when using synonyms, and 1,643 when using both synonyms and hypernyms, accounting for 43% and 66% of the test sentences, respectively.

The number of alternative sentences generated for each source text varies from 1 to 960 when paraphrasing rules were applied, and reaches very large numbers, up to 89,700 at the "worst case", when all TE rules are employed, an average of 456 alternatives per sentence.

Scoring source texts We test our proposed method using several context-models shown to perform reasonably well in previous work:

- ***FREQ***: The first model we use is a context-independent baseline. A common useful heuristic to pick an entailment rule is to select the candidate with the highest frequency in the corpus (Mccarthy et al., 2004). In this model, a rule's score is the normalized number of occurrences of its RHS in the training corpus, ignoring the context of the LHS.
- ***LSA***: Latent Semantic Analysis (Deerwester et al., 1990) is a well-known method for representing the contextual usage of words based on corpus statistics. We represented each term by a normalized vector of the top 100 SVD dimensions, as described in (Gliozzo, 2005). This model measures the similarity between the sentence words and the RHS in the LSA space.
- ***NB***: We implemented the unsupervised Naïve Bayes model described in (Glickman et al., 2006) to estimate the probability that the unknown term entails the RHS in the given context. The estimation is based on corpus co-occurrence statistics of the context words with the RHS.
- ***LMS***: This model generates the Language Model probability of the RHS in the source. We use 3-grams probabilities as produced by the SRILM toolkit (Stolcke, 2002).

Finally, as a simple baseline, we generated a random score for each rule application, ***RAND***.

The score of each rule application by any of the above models is normalized to the range (0,1]. To combine individual rule applications in a given sentence, we use the product of their scores. The monolingual data used for the models above is the source side of the training parallel corpus.

Target-language scores On the target side we used either a standard 3-gram language-model, denoted ***LMT***, or the score assigned by the complete SMT log-linear model, which includes the language model as one of its components (***SMT***).

A pair of a *source*:*target* models comprises a complete model for selecting the best translated sentence, where the overall score is the product of the scores of the two models.

We also applied several combinations of source models, such as *LSA* combined with *LMS*, to take advantage of their complementary strengths. Additionally, we assessed our method with source-only models, by setting the number of sentences to be selected by the source model to one ($k = 1$).

5 Results

5.1 Manual Evaluation

To evaluate the translations produced using the various source and target models and the different rule-sets, we rely mostly on manual assessment, since automatic MT evaluation metrics like BLEU do not capture well the type of semantic variations

[2] http://opennlp.sourceforge.net

	Model	Precision (%)		Coverage (%)	
		PARAPH.	TE	PARAPH.	TE
1	*–:SMT*	75.8	73.1	32.5	48.1
2	*NB:SMT*	75.2	71.5	32.3	47.1
3	*LSA:SMT*	74.9	72.4	32.1	47.7
4	*NB:–*	74.7	71.1	32.1	46.8
5	*LMS:LMT*	73.8	70.2	31.7	46.3
6	*FREQ:–*	72.5	68.0	31.2	44.8
7	*RAND*	57.2	63.4	24.6	41.8

Table 1: Translation acceptance when using only paraphrases and when using all entailment rules. "*:*" indicates which model is applied to the source (left side) and which to the target language (right side).

generated in our experiments, particularly at the sentence level.

In the manual evaluation, two native speakers of the target language judged whether each translation preserves the meaning of its reference sentence, marking it as acceptable or unacceptable. From the sentences for which rules were applicable, we randomly selected a sample of sentences for each annotator, allowing for some overlapping for agreement analysis. In total, the translations of 1,014 unique source sentences were manually annotated, of which 453 were produced using only hypernyms (no paraphrases were applicable). When a sentence was annotated by both annotators, one annotation was picked randomly.

Inter-annotator agreement was measured by the percentage of sentences the annotators agreed on, as well as via the Kappa measure (Cohen, 1960). For different models, the agreement rate varied from 67% to 78% (72% overall), and the Kappa value ranged from 0.34 to 0.55, which is comparable to figures reported for other standard SMT evaluation metrics (Callison-Burch et al., 2008).

Translation with TE For each model m, we measured $Precision_m$, the percentage of acceptable translations out of all sampled translations. $Precision_m$ was measured both when using only paraphrases (PARAPH.) and when using all entailment rules (TE). We also measured $Coverage_m$, the percentage of sentences with acceptable translations, A_m, out of all sentences (2,494). As our annotators evaluated only a sample of sentences, A_m is estimated as the model's total number of sentences with applicable rules, S_m, multiplied by the model's Precision (S_m was 1,071 for paraphrases and 1,643 for entailment rules): $Coverage_m = \frac{S_m \cdot Precision_m}{2{,}494}$.

Table 1 presents the results of several source-target combinations when using only paraphrases and when also using directional entailment rules. When all rules are used, a substantial improvement in coverage is consistently obtained across all models, reaching a relative increase of 50% over paraphrases only, while just a slight decrease in precision is observed (see Section 5.3 for some error analysis). This confirms our hypothesis that directional entailment rules can be very useful for replacing unknown terms.

For the combination of source-target models, the value of *k* is set depending on which rule-set is used. Preliminary analysis showed that $k = 5$ is sufficient when only paraphrases are used and $k = 20$ when directional entailment rules are also considered.

We measured statistical significance between different models for precision of the TE results according to the Wilcoxon signed ranks test (Wilcoxon, 1945). Models 1-6 in Table 1 are significantly better than the *RAND* baseline ($p < 0.03$), and models 1-3 are significantly better than model 6 ($p < 0.05$). The difference between *–:SMT* and *NB:SMT* or *LSA:SMT* is not statistically significant.

The results in Table 1 therefore suggest that taking a source model into account preserves the quality of translation. Furthermore, the quality is maintained even when source models' selections are restricted to a rather small top-k ranks, at a lower computational cost (for the models combining source and target, like *NB:SMT* or *LSA:SMT*). This is particularly relevant for on-demand MT systems, where time is an issue. For such systems, using this source-language based pruning methodology will yield significant performance gains as compared to target-only models.

We also evaluated the baseline strategy where unknown terms are omitted from the translation, resulting in 25% precision. Leaving unknown words untranslated also yielded very poor translation quality in an analysis performed on a similar dataset.

Comparison to related work We compared our algorithm with an implementation of the algorithm proposed by (Callison-Burch et al., 2006) (see Section 2.2), henceforth *CB*, using the Spanish side of Europarl as the pivot language.

Out of the tested 2,494 sentences with unknown terms, *CB* found paraphrases for 706 sentences (28.3%), while with any of our models, including

Model	Precision (%)	Coverage (%)	Better (%)
NB:SMT (TE)	85.3	56.2	72.7
CB	85.3	24.2	12.7

Table 2: Comparison between our top model and the method by Callison-Burch et al. (2006), showing the percentage of times translations were considered acceptable, the model's coverage and the percentage of times each model scored better than the other (in the 14% remaining cases, both models produced unacceptable translations).

Model	BLEU (TE)	METEOR (TE)
–:SMT	15.50	0.1325
NB:SMT	15.37	0.1316
LSA:SMT	15.51	0.1318
NB:–	15.37	0.1311
CB	15.33	0.1299
Baseline SMT	15.29	0.1294

Table 3: Performance of the best models according to automatic MT evaluation metrics at the corpus level. The baseline refers to translation of the text without applying any entailment rules.

NB:SMT, our algorithm found applicable entailment rules for 1,643 sentences (66%).

The quality of the *CB* translations was manually assessed for a sample of 150 sentences. Table 2 presents the precision and coverage on this sample for both *CB* and *NB:SMT*, as well as the number of times each model's translation was preferred by the annotators. While both models achieve equally high precision scores on this sample, the *NB:SMT* model's translations were undoubtedly preferred by the annotators, with a considerably higher coverage.

With the *CB* method, given that many of the phrases added to the phrase table are noisy, the global quality of the sentences seem to have been affected, explaining why the judges preferred the *NB:SMT* translations. One reason for the lower coverage of *CB* is the fact that paraphrases were acquired from a corpus whose domain is different from that of the test sentences. The entailment rules in our models are not limited to paraphrases and are derived from WordNet, which has broader applicability. Hence, utilizing monolingual resources has proven beneficial for the task.

5.2 Automatic MT Evaluation

Although automatic MT evaluation metrics are less appropriate for capturing the variations generated by our method, to ensure that there was no degradation in the system-level scores according to such metrics we also measured the models' performance using BLEU and METEOR (Agarwal and Lavie, 2007). The version of METEOR we used on the target language (French) considers the stems of the words, instead of surface forms only, but does not make use of WordNet synonyms.

We evaluated the performance of the top models of Table 1, as well as of a baseline SMT system that left unknown terms untranslated, on the sample of 1,014 manually annotated sentences. As shown in Table 3, all models resulted in improvement with respect to the original sentences (baseline). The difference in METEOR scores is statistically significant ($p < 0.05$) for the three top models against the baseline. The generally low scores may be attributed to the fact that training and test sentences are from different domains.

5.3 Discussion

The use of entailed texts produced using our approach clearly improves the quality of translations, as compared to leaving unknown terms untranslated or omitting them altogether. While it is clear that textual entailment is useful for increasing coverage in translation, further research is required to identify the amount of *information loss* incurred when non-symmetric entailment relations are being used, and thus to identify the cases where such relations are detrimental to translation.

Consider, for example, the sentence: "*Conventional military models are geared to **decapitate** something that, in this case, has no head.*". In this sentence, the unknown term was replaced by *kill*, which results in missing the point originally conveyed in the text. Accordingly, the produced translation does not preserve the meaning of the source, and was considered unacceptable: "*Les modèles militaires visent à **faire** quelque chose que, dans ce cas, n'est pas responsable.*".

In other cases, the selected hypernyms were too generic words, such as *entity* or *attribute*, which also fail to preserve the sentence's meaning. On the other hand, when the unknown term was a very specific word, hypernyms played an important role. For example, "*Bulgaria is the most sought-after east European real estate target, with its low-cost ski **chalets** and oceanfront homes*". Here, *chalets* are replaced by *houses* or *units* (depending on the model), providing a translation that would be acceptable by most readers.

Other incorrect translations occurred when the unknown term was part of a phrase, for example, *troughs* replaced with *depressions* in *peaks*

and troughs, a problem that also strongly affects paraphrasing. In another case, *movement* was the hypernym chosen to replace *labor* in *labor movement*, yielding an awkward text for translation.

Many of the cases which involved ambiguity were resolved by the applied context-models, and can be further addressed, together with the above mentioned problems, with better source-language context models.

We suggest that other types of entailment rules could be useful for the task beyond the straightforward generalization using hypernyms, which was demonstrated in this work. This includes other types of lexical entailment relations, such as holonymy (e.g. *Singapore* ⇒ *Southeast Asia*) as well as lexical syntactic rules (*X cure Y* ⇒ *treat Y with X*). Even syntactic rules, such as clause removal, can be recruited for the task: *"Obama, the 44th president, declared Monday . . . "* ⇒ *"Obama declared Monday . . . "*. When the system is unable to translate a term found in the embedded clause, the translation of the less informative sentence may still be acceptable by readers.

6 Conclusions and Future Work

In this paper we propose a new entailment-based approach for addressing the problem of unknown terms in machine translation. Applying this approach with lexical entailment rules from WordNet, we show that using monolingual resources and textual entailment relationships allows substantially increasing the quality of translations produced by an SMT system. Our experiments also show that it is possible to perform the process efficiently by relying on source language context-models as a filter prior to translation. This pipeline maintains translation quality, as assessed by both human annotators and standard automatic measures.

For future work we suggest generating entailed texts with a more extensive set of rules, in particular lexical-syntactic ones. Combining rules from monolingual and bilingual resources seems appealing as well. Developing better context-models to be applied on the source is expected to further improve our method's performance. Specifically, we suggest taking into account the prior likelihood that a rule is correct as part of the model score.

Finally, some researchers have advocated recently the use of shared structures such as parse forests (Mi and Huang, 2008) or word lattices (Dyer et al., 2008) in order to allow a compact representation of alternative inputs to an SMT system. This is an approach that we intend to explore in future work, as a way to efficiently handle the different source language alternatives generated by entailment rules. However, since most current MT systems do not accept such type of inputs, we consider the results on pruning by source-side context models as broadly relevant.

Acknowledgments

This work was supported in part by the ICT Programme of the European Community, under the PASCAL 2 Network of Excellence, ICT-216886 and The Israel Science Foundation (grant No. 1112/08). We wish to thank Roy Bar-Haim and the anonymous reviewers of this paper for their useful feedback. This publication only reflects the authors' views.

References

Abhaya Agarwal and Alon Lavie. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of WMT-08*.

Karunesh Arora, Michael Paul, and Eiichiro Sumita. 2008. Translation of Unknown Words in Phrase-Based Statistical Machine Translation for Languages of Rich Morphology. In *Proceedings of SLTU*.

Francis Bond, Eric Nichols, Darren Scott Appling, and Michael Paul. 2008. Improving Statistical Machine Translation by Paraphrasing the Training Data. In *Proceedings of IWSLT*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of HLT-NAACL*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceedings of WMT*.

Chris Callison-Burch. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. In *Proceedings of EMNLP*.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Trevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of ACL*.

Ido Dagan, Oren Glickman, Alfio Massimiliano Gliozzo, Efrat Marmorshtein, and Carlo Strapparava. 2006. Direct Word Sense Matching for Lexical Substitution. In *Proceedings of ACL*.

Scott Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R.A. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing Word Lattice Translation. In *Proceedings of ACL-HLT*.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2008. Communicating Unknown Words in Machine Translation. In *Proceedings of LREC*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The Third PASCAL Recognising Textual Entailment Challenge. In *Proceedings of ACL-WTEP Workshop*.

Oren Glickman, Ido Dagan, Mikaela Keller, Samy Bengio, and Walter Daelemans. 2006. Investigating Lexical Substitution Scoring for Subtitle Generation. In *Proceedings of CoNLL*.

Alfio Massimiliano Gliozzo. 2005. *Semantic Domains in Computational Linguistics*. Ph.D. thesis, University of Trento.

Francisco Guzmán and Leonardo Garrido. 2008. Translation Paraphrases in Phrase-Based Machine Translation. In *Proceedings of CICLing*.

Nizar Habash. 2008. Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation. In *Proceedings of ACL-HLT*.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of HLT-NAACL*.

Kevin Knight and Jonathan Graehl. 1997. Machine Transliteration. In *Proceedings of ACL*.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of EACL*.

Philippe Langlais and Alexandre Patry. 2007. Translating Unknown Words by Analogical Learning. In *Proceedings of EMNLP-CoNLL*.

Dekang Lin and Patrick Pantel. 2001. DIRT – Discovery of Inference Rules from Text. In *Proceedings of SIGKDD*.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of SemEval*.

Diana Mccarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding Predominant Word Senses in Untagged Text. In *Proceedings of ACL*.

Haitao Mi and Liang Huang. 2008. Forest-based Translation Rule Extraction. In *Proceedings of EMNLP*.

Sebastian Pado, Michel Galley, Daniel Jurafsky, and Christopher D. Manning. 2009. Textual Entailment Features for Machine Translation Evaluation. In *Proceedings of WMT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*.

M. Simard, N. Cancedda, B. Cavestro, M. Dymetman, E. Gaussier, C. Goutte, and K. Yamada. 2005. Translating with Non-contiguous Phrases. In *Proceedings of HLT-EMNLP*.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of ICSLP*.

Idan Szpektor, Ido Dagan, Roy Bar-Haim, and Jacob Goldberger. 2008. Contextual Preferences. In *Proceedings of ACL-HLT*.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages. In *Proceedings of EACL*.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2008. Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora. In *Proceedings of ACL-HLT*.

Chapter 5

Classification-based Contextual Preferences

Classification-based Contextual Preferences

Shachar Mirkin, Ido Dagan, Lili Kotlerman
Bar-Ilan University
Ramat Gan, Israel
{mirkins,dagan,davidol}@cs.biu.ac.il

Idan Szpektor
Yahoo! Research
Haifa, Israel
idan@yahoo-inc.com

Abstract

This paper addresses context matching in textual inference. We formulate the task under the *Contextual Preferences* framework which broadly captures contextual aspects of inference. We propose a generic classification-based scheme under this framework which coherently attends to context matching in inference and may be employed in any inference-based task. As a test bed for our scheme we use the Name-based Text Categorization (TC) task. We define an integration of Contextual Preferences into the TC setting and present a concrete self-supervised model which instantiates the generic scheme and is applied to address context matching in the TC task. Experiments on standard TC datasets show that our approach outperforms the state of the art in context modeling for Name-based TC.

1 Introduction

Textual inference is prevalent in text understanding applications. For example, in Question Answering (QA) the expected answer should be inferred from retrieved passages, and in Information Extraction (IE) the meaning of the target event is inferred from its mention in the text.

Lexical inferences make a substantial part of the inference process. In such cases, a target term is inferred from text expressions based on either one of two types of lexical matches: (i) a *direct match* of the target term in the text. For instance, the IE event injure may be detected by finding the word *injure* in the text; (ii) an *indirect match*, through a term that implies the meaning of the target term, e.g. inferring *injure* from *hurt*.

In either case, due to word ambiguity, it is necessary to validate that the context of the match conforms with the intended meaning of the target term before carrying out an inference operation based on this match. For example, "*You hurt my feelings*" constitutes an invalid context for the injure event as *hurt* in this text does not refer to a physical injury. Similarly, inferring the protest-related event demonstrate based on *demo* is deemed invalid although *demo* implies the meaning of the word *demonstrate* in other contexts, e.g., concerning *software demonstration*.

Although seemingly equivalent, a closer look reveals that the above two examples correspond to two distinct contextual mismatch situations. While the match of *hurt* is invalid for injure in the particular given context, an inference based on *demo* is invalid for the protest demonstrate event in any context.

Thus, several types of context matching are involved in textual inference. While most prior work addressed only specific context matching scenarios, Szpektor et al. (2008) presented a broader view, proposing a generic framework for context matching in inference, termed Contextual Preferences (CP). CP specifies the types of context matching that need to be considered in inference, allowing a model of choice to be applied for validating each type of match. Szpektor et al. applied CP to an IE task using different models to validate each type of context match.

In this work we adopt CP as our context matching framework and propose a novel classification-based scheme which provides unified modeling for CP. We represent typical contexts of the textual objects that participate in inference using classifiers; at inference time, each match is assessed by the respective classifiers which determine its contextual validity.

As a test bed we applied our scheme to the task

Proceedings of the TextInfer 2011 Workshop on Textual Entailment, EMNLP 2011, pages 20–29,
Edinburgh, Scotland, UK, July 30, 2011.

of Name-based Text Categorization. This is an unsupervised setting of TC where the only input given is the category name, and in which context validation is of high importance. We instantiate the scheme with a novel self-supervised model and apply it to the TC task. We suggest a method for integrating any CP-based context matching model into TC and use it to combine the context matching scores generated by our model. Results on two standard TC datasets show that our approach outperforms the state of the art context model for this task and suggest applying this scheme to additional inference-based applications.

2 Background

2.1 Context matching in inference

Word ambiguity has been traditionally addressed through Word Sense Disambiguation (WSD) (Navigli, 2009). The WSD task requires selecting the meaning of a target term from amongst a predefined set of senses, based on sense-inventories such as WordNet (Fellbaum, 1998).

An alternative approach eliminates the reliance on such inventories. Instead of explicit sense identification, a direct sense-match between terms is pursued (Dagan et al., 2006). *Lexical substitution* (McCarthy and Navigli, 2009) is probably the most commonly known task that follows this approach. *Context matching* is a generalization of lexical substitution, which seeks a match between terms in context, not necessarily for the purpose of substitution. For instance, the word *played* in "*U2 played their first-ever concert in Russia*" contextually matches *music*, although *music* cannot substitute *played* in this context. The context matching task, therefore, is to determine (by quantifying or giving a binary decision) the validity of a match between two terms in context.

In Section 1 we informally presented two cases of contextual mismatches. A comprehensive view of context matching types is provided by the Contextual Preferences framework (Szpektor et al., 2008). CP is phrased in terms of the Textual Entailment (TE) paradigm (Dagan et al., 2009). In TE, a *text* t *entails* a textual *hypothesis* h if the meaning of h can be inferred from t. Formulating the IE example from Section 1 within TE, h may be the name of the target event, *injure*, and t is a text segment from which h can be inferred. A direct match occurs when a term in h is identical to a term in t. An inference based on an indirect match is viewed as the application of a *lexical entailment rule*, r, such as '*hurt* $\Rightarrow$ *injure*', where the entailing left-hand side (LHS) of the rule (*hurt*) is matched in the text, while the entailed right-hand side (RHS), *injure*, is matched in the hypothesis.

Hence, three *inference objects* take part in inference operations: t, h and r. Most prior work addressed only specific contextual matches between these objects. For example, Harabagiu et al. (2003) matched the contexts of t and h for QA (answer and question, respectively); Barak et al. (2009) matched t and h (document and category) in TC, while other works, including those applying lexical substitution, typically validated the context match between t and r (Kauchak and Barzilay, 2006; Dagan et al., 2006; Pantel et al., 2007; Connor and Roth, 2007).

In comparison, in the CP framework, all possible contextual matches among t, h and r are considered: $t-h$, $t-r$ and $r-h$. The three context matches are depicted in Figure 1 (left). In CP, the representation of each inference object is enriched with contextual information which is used to characterize its valid contexts. Such information may be the words of the event description in IE, corpus instances based on which a rule was learned, or an annotation of relevant WordNet senses in Name-based TC. For example, a category name hockey may be assigned with the sense number corresponding to *ice hockey*, but not to *field hockey*, in order to designate information that limits the valid contexts of the category to the former among the two meanings of the name.

Before an inference operation is performed, the context representations of each pair among the participating objects should be *matched* by a context model in order to assess the contextual validity of the operation. Along with the context representation and the specific context matching models, the way context model decisions are combined needs to be specified in a concrete implementation of the CP framework.

2.2 Context matching models

Several approaches were taken in prior work to model context matching, mostly within the scope of learning *selectional preferences* of templatized lexical-syntactic rules (e.g. '$X \xleftarrow{subj}$ *hit* $\xrightarrow{obj} Y$' $\Rightarrow$ '$X \xleftarrow{subj}$ *attack* $\xrightarrow{obj} Y$').

Pantel et al. (2007) and Szpektor et al. (2008) represented the context of such rules as the intersection of preferences of the rule's LHS and RHS, namely the observed argument instantiations or their semantic classes. A rule is deemed applicable to a given text if the argument instantiations in the text are similar to the selectional preferences of the rule. To overcome sparseness, other works represented context in latent space. Pennacchiotti et al. (2007) and Szpektor et al. (2008) measured the similarity between the Latent Semantic Analysis (LSA) (Deerwester et al., 1990) representations of matched contexts. Dinu and Lapata (2010) used Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to model templates' latent senses, determining rule applicability based on the similarity between the two sides of the rule when instantiated by the context, while Ritter et al. (2010) used LDA to model argument classes, considering a rule valid for a given argument instantiation if its instantiated templates are drawn from the same hidden topic.

A different approach is provided by classification-based models which learn classifiers for inference objects. A classifier is trained based on positive and negative examples which represent valid or invalid contexts of the object; from those, features characterizing the context are extracted, e.g. words in a window around the target term or syntactic links with it. Given a new context, the classifier assesses its validity with respect to the learned classification model.

Classifiers in prior work were applied to determine rule applicability in a given context ($t - r$). Training a classifier for word paraphrasing, Kauchak and Barzilay (2006) used occurrences of the rule's RHS as positive context examples, and randomly picked negative examples. A similar approach was applied by Dagan et al. (2006), which used a single-class SVM to avoid selecting negative examples. In both works, a resulting classifier represents a word with all its senses intermixed. Clearly, this poses no problem for monosemous words, but is biased towards the more common senses of polysemous words. Indeed, Dagan et al. (2006) report a negative correlation between the degree of polysemy of a word and the performance of its classifier. Connor and Roth (2007) used per-rule classifiers to produce a noisy training set for learning a global classifier for verb substitution.

In this work we follow the classification-based approach which seems appealing for several reasons.

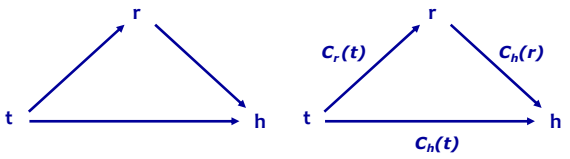

Figure 1: Left: An illustration of the CP relationships as in (Szpektor et al., 2008), with arrows indicting the context matching direction; Right: The application of classifiers to the tested contexts under our scheme.

First, it allows seamlessly integrating various types of information via classifiers' features; unlike some of the above models, it is not inherently dependent on the type of rules that are utilized and easily accommodates to both lexical and lexical-syntactic rules through the choice of features. In addition, it does not rely on a predefined similarity measure and provides flexibility in terms of model's parameters. Finally, this approach captures the notion of directionality which is fundamental in textual inference, and is therefore better suited to applied inference than previously proposed symmetric context models.

In comparison to prior classification-based models, our approach addresses all three context matches specified by CP, rather than only the rule-text match. It is not limited to substitutable terms or even to terms with the same part of speech. In addition, we avoid learning a classifier for all senses combined, but rather learn it for the specific intended meaning.

2.3 Name-based Text Categorization

Name-based TC (Gliozzo et al., 2009) is an unsupervised setting of Text Categorization in which the only input provided is the category name, e.g. trade, 'mergers and acquisitions' or guns. When category names are ambiguous, e.g. space, categories are not well defined; thus, auxiliary information is expected to accompany the name for disambiguation, such as a list of relevant senses or a category description.

Typically, unsupervised TC consists of two steps. First, an unsupervised method is applied to an unlabeled corpus, automatically labeling some of the documents to categories. Then, the labeled documents from the first step are used to train a supervised TC classifier which is used to label any document in the test set (Gliozzo et al., 2009; Downey and Etzioni, 2009; Barak et al., 2009).

In this work we focus on the above unsupervised step. Gliozzo et al. (2009) addressed this task by representing both documents and categories by LSA vectors which implicitly capture contextual similarities between terms. Each document was then assigned to the most similar category based on cosine similarity between the LSA vectors. Barak et al. (2009) required an occurrence of a term entailing the category name (or the category name itself) in order to regard the category as a candidate for the document. To assess the contextual validity of the match, they used LSA document-category similarity as in (Gliozzo et al., 2009). For example, to classify a document into the category medicine, at least one lexical entailment rule, e.g. '*drug* ⇒ *medicine*', should be matched in the document. Then, the validity of *drug* for medicine in the matched document is assessed by the LSA context model. In this work we adopt Barak et al.'s requirement for a match for the category in the document, but address context matching in an entirely different way.

Name-based TC provides a convenient setting for evaluating context matching approaches for two main reasons. First, all types of context matchings are realized in this application (see Section 3); second, as the hypothesis consists of a single term or a few terms, the TC gold standard annotation corresponds quite directly to the context matching task for lexical inferences; in other applications where longer hypotheses are involved, context matching performance may be masked by other factors.

3 Contextual Matches in TC

Within Name-based TC, the Textual Entailment terminology is mapped as follows: h is a term denoting the category name (e.g. *merger* or *acquisition*); t is a matched term in the document to be categorized from which h may be inferred; and a match refers to an occurrence in the document of either h (direct match) or the LHS of an entailment rule r whose RHS is a category name (indirect match).[1]

Under the CP view, a context model needs to address the following three context matching cases within a TC setting.

$t-h$: Assessing the validity of a match in the document with respect to the category's intended meaning. For example, the occurrence of the category name space (in the sense of *outer space*) in "*the server ran out of disk space*" does not indicate a space-related text, and should be dismissed by the context model.

$t - r$: This case refers to a rule match in the document. A context model should ensure that the meaning of a match is compatible with that of the rule. For example, '*alien* ⇒ *space*' is a valid rule for the space category. Yet, it should not be applied to "*The US welcomes a large number of aliens every year*", since *alien* in this sentence has a different meaning than the intended meaning of the rule.

$r - h$: The match between the intended meanings of the category name and the RHS of the rule. For instance, the rule '*room* ⇒ *space*' is not suitable at all for the (outer) space category.

4 A Classification-based Scheme for CP

Szpektor et al. (2008) introduced a vector-space model to implement CP, in which the text t, the rule r and the hypothesis h share the same contextual representation. However, in CP, r, h and t have non-symmetric roles: the context of t should be tested as valid for r and h and not vice versa, and the context of r should be validated for h and not the other way around. This stems from the need to consider directionality in context matching. For instance, a text about *football* typically constitutes a valid context for the more general *sports* context, but not vice versa. Indeed, directionality may be captured in vector-space models by using a directional similarity measure (Kotlerman et al., 2010), but only symmetric measures were used in context matching work so far.

Based on this distinction between the inference objects' roles, we present a novel scheme that uses two types of classifiers to represent context:

C_h: A classifier that identifies valid contexts for h. It tests contexts of t (for $t - h$ matching) or r (for $r - h$ matching), assigning them scores $C_h(t)$ and $C_h(r)$, respectively.

C_r: A classifier that identifies valid contexts for applying the rule r. It tests the context of t, assigning it a score $C_r(t)$.

Figure 1 (right) shows the classifiers scores which are assigned to each of the matching types.

[1] Note that t and h both refer here to individual terms.

Hence, h always acts as the *classifying object*, t is always the *classified object*, while r acts as both. Context matching is quantified by the degree by which the classified object represents a valid context for the classifying object in a given inference scenario.

In comparison to the CP implementation in (Szpektor et al., 2008), our approach uses a unified model which captures directionality in context matching.

To instantiate the scheme, one needs to define the way training examples are obtained and processed. This may be done within supervised classification, where labeled examples are provided, or – as we do in this work – using self-supervised classifiers which obtain training examples automatically. We present such an instantiation in Section 5, where a classifier is trained for each category and each rule. When more complex hypotheses are involved, C_h classifiers can be trained separately for each relevant part of the hypothesis, using the rest for disambiguation.

A combination of the three model scores provides a final context matching score. In Section 6 we suggest a way to combine the actual classification scores as part of the integration of CP into TC, but other combinations are plausible. In particular, binary classifications (valid vs. invalid) may be used as filters. That is, the context is classified as valid only if all relevant models classify it as such.

5 A Self-supervised Context Model

We now turn to demonstrate how our classification-based scheme may be implemented. The model below is exemplified on Name-based TC, but may be applicable to other tasks, with few changes.

5.1 Training-set generation

Our implementation is self-supervised as we want to integrate it within the unsupervised TC setting. That is, the classifiers automatically obtain training examples for the classifying object (a category or a rule) without relying on labeled documents.

We obtain examples by querying the TC training corpus with automatically-generated search queries. The difficulty lies in correctly constructing queries that will retrieve documents representing either valid or invalid contexts for the classifying object. To this end, we retrieve examples through a gradual process in which the most accurate (least ambiguous) query is used first and less accurate queries follow, until the designated number of examples is acquired.

5.1.1 Obtaining positive examples

To acquire positive training examples, we construct queries which are comprised of two main clauses. The first contains the *seeds*, terms which characterize the classifying object. Primarily, these are the category name or the LHS of the rule. The second consists of *context words* which are used when the seeds are polysemous, and are intended to assist disambiguation. When context words are used, at least one seed and at least one context word must be matched to retrieve a document. For example, given the highly ambiguous category name space, we first construct the query using only the monosemous term *outer space*; if the number of retrieved documents does not meet the requirement, a second query may be constructed: *("outer space"* OR *space)* AND *(infinite* OR *science* OR ... *)*.

To generate a rule classifier C_r, we retrieve positive examples as follows. If the LHS term is monosemous according to WordNet[2], we first query using this term alone (e.g. *decrypt*), and add its monosemous synonyms and hyponyms if more examples are required (e.g. *decrypt* OR *decode*). If the LHS is polysemous, we carry out Procedure 1. Intuitively, this procedure tries to minimize ambiguity by using monosemous terms as much as possible; when polysemous terms must be used, it tries to ensure there are monosemous terms to disambiguate them. Note that entailment directionality is maintained throughout the process, as seeds are only expanded with more specific (entailing) terms, while context words are only expanded with more general (entailed) terms.

Procedure 1 : Retrieval of C_r positive examples

Apply sequentially until sufficient examples are obtained:

1: Set the LHS as seed and the RHS's monosemous synonyms, hypernyms and derivations as context words.
2: Add monosemous synonyms and hyponyms of the LHS to the seeds.
3: As in 2, but use polysemous terms as well.
4: Add polysemous context words.

Positive examples for category classifiers (C_h) are obtained through a similar procedure as for rule clas-

[2]Terms not in WordNet are assumed monosemous.

sifiers. If the category is part of a hierarchy, we also use the name of the parent category (e.g. *sport* for rec.sport.hockey) as a context word.

5.1.2 Obtaining negative examples

Negative examples are even more challenging to acquire. In prior work negative examples were selected randomly (Kauchak and Barzilay, 2006; Connor and Roth, 2007). We follow this method, but also attempt to identify negative examples that are semantically similar to the positive ones in order to improve the discriminative power of the classifier (Smith and Eisner, 2005). We do that by applying a similar procedure which uses cohyponyms of the seeds, e.g. *baseball* for hockey or *islam* for christianity. Cohyponymy is a non-entailing relation; hence, by using it we expect to obtain semantically-related, yet invalid contexts. If not enough negative examples are retrieved using cohyponyms, we select the remaining required examples randomly.

As the distribution of positive and negative examples in the data is unknown, we set the ratio of negative to positive examples as a parameter of the model, as in (Bergsma et al., 2008).

5.1.3 Insufficient examples

When the number of training examples for a rule or a category is below a certain minimum, the resulting classifier is expected to be of poor quality. This usually happens for positive examples in any of the following two cases: (i) the seed is rare in the training set; (ii) the desired sense of the seed is rarely found in the training set, and unwanted senses were filtered by our retrieval query. For instance, *nazarene* does not occur at all in the training set, and the classifier corresponding to the rule '*nazarene* $\Rightarrow$ *christian*' cannot be generated. On the other hand, *cone* does appear in the corpus but not in the astrophysical sense the rule '*cone* $\Rightarrow$ *space*' refers to. In such cases we refrain from generating the classifier and use instead a default score of 0 for each classified object. The idea is that rare terms will also occur infrequently in the test set, while cases where the term is found in the corpus, but in a different sense than the desired one, will be blocked.

5.1.4 Feature extraction

We extract global and local lexical features that are standard in WSD work. Global features include all the terms in the document or in the sentence in which a match was found. Local features are extracted around matches of seeds which comprised the query that retrieved the document. These features include the terms in a window around the match, and the noun, verb, adjective and adverb nearest to the match in either direction. For randomly sampled negative examples, where no matched query terms exist, we randomly select terms in the document as "matches" for local feature extraction. If more than one match of the same term is found in a document, we assume one-sense-per-discourse (Gale et al., 1992) and jointly extract features for all matches of the term.

5.2 Applying the classifiers

During inference, for each direct match in a document, the corresponding C_h is applied. For an indirect match, the respective C_r is also applied.

In addition, C_h is applied to the matched rules. Unlike t, a rule is not represented by a single text. Therefore, to test a rule's match with the category, we randomly sample from the training set documents containing the rule's LHS. We apply C_h to each sampled example and compute the ratio of positive classifications. The result is a score indicating the domain-specific probability of the rule to be applicable to the category, and may be interpreted as an in-domain *prior*. For instance, the rule '*check* $\Rightarrow$ *hockey*' is assigned a score of 0.05, since the sense of *check* as a hockey defense technique is rare in the corpus. On the other hand, non ambiguous rules, e.g. '*warship* $\Rightarrow$ *ship*' are assigned a high probability (1.0), and so are rules whose LHS is ambiguous but its dominant sense in the training corpus is the same one the rule refers to, e.g. '*margin* $\Rightarrow$ *earnings*'(0.85).

We do not assign negative classifier scores to invalid matches but rather set them to zero instead. The reason is that an invalid context only indicates that the term cannot be used for entailing the category name, but not that the document itself is irrelevant.

6 CP for Text Categorization

CP may be employed in any inference-based task, but the integration with each task is somewhat different and needs to be specified. Below we present a methodology for integrating CP into Name-based Text Categorization.

As in (Barak et al., 2009) (*Barak09* below), we represent documents and categories by term-vectors in the following way: a document vector contains the document terms; a category vector contains two sets of terms: $\mathcal{C}$, the terms denoting the category name, and $\mathcal{E}$, their entailing terms. For example, *oil* is added to the vector of the category crude by the rule '*oil* $\Rightarrow$ *crude*' (i.e. *crude* $\in \mathcal{C}$ and *oil* $\in \mathcal{E}$).

Barak09 assigned equal values of 1 to all vector entries. We suggest integrating a CP-based context model into TC by re-weighting the terms in the vectors, prior to determining the final document-category categorization score through vector similarity. Given a category c, with term vector C, and a document d with term vector D, the model re-weights vector entries of matching terms (i.e., terms in $C \cap D$), based on the validity of the context match. Valid matches should be assigned with higher scores than invalid ones, leading to higher overall vector similarity for documents with valid matches for the given category. Non-matching terms are ignored as their weights are canceled out in the subsequent vector product.

Specifically, the model assigns a new weight $w_D(u)$ to a matching term u in the document vector D based on the model's assessment of: (a) $t-h$, the context match between the (match in the) document and the category; and (if an indirect match) (b) $t-r$, the context match between the document and the rule '$u \Rightarrow c_i$', where $c_i \in \mathcal{C}$. The model also sets a new weight $w_C(v)$ to a term v in the category vector C based on the context match for $r-h$, between the rule '$v \Rightarrow c_j$' ($c_j \in \mathcal{C}$) and the category. For instance, using our context matching scheme in TC, $w_D(u)$ is set to $C_h(u)$ or $\frac{C_h(u)+C_r(u)}{2}$ for direct and indirect matches, respectively; $w_C(v)$ is left as 1 if $v \in \mathcal{C}$ and set to $C_h(v)$ when $v \in \mathcal{E}$.

Barak09 assigned a single global context score to a document-category pair using the LSA representations of their vectors. In our approach, however, we consider the actual matches from the three different views, hence the re-weighting of the vector entries using three model scores.

7 Experimental Setting

7.1 Datasets and knowledge resources

Following (Gliozzo et al., 2009) and (Barak et al., 2009), we evaluated our method on two standard TC datasets: *Reuters-10* and *20-Newsgroups*.

The *Reuters-10* (R10, for short) is a sub-corpus of the Reuters-21578 collection[3], constructed from the ten most frequent categories in the Reuters taxonomy. We used the Apte split of the Reuters-21578 collection, often used in TC tasks. The top 10 categories include about 9,000 documents, split into training (70%) and test (30%) sets. The *20-Newsgroups* (20NG) corpus is a collection of newsgroup postings gathered from twenty different categories from the Usenet Newsgroups hierarchy[4]. We used the "by-date" version of the corpus, which contains approximately 20,000 documents partitioned (nearly) evenly across the categories and divided in advance to training (60%) and test (40%) sets.

As in (Gliozzo et al., 2009; Barak et al., 2009), we adjusted non-standard category names (e.g. forsale was renamed to sale) and manually specified for each category its relevant WordNet senses. The sense tagging properly defines the categories, and is expected to accompany such hypotheses. Other types of information may be used for this purpose, e.g. words from category descriptions, if such exist.

We applied standard preprocessing (sentence splitting, tokenization, lemmatization and part of speech tagging) to all documents in the datasets. All terms, including those denoting category names and rules, are represented by their lemma and part of speech.

As sources for lexical entailment rules we used WordNet 3.0 (synonyms, hyponyms, derivations and meronyms) and a Wikipedia-derived rule-base (Shnarch et al., 2009). Unlike *Barak09* we did not limit the rules extracted from WordNet to the most frequent senses and used all rule types from the Wikipedia-based resource.

7.2 Self-supervised model tuning

Tuning of the self-supervised context model's parameters (number of training examples, negative to positive ratio, feature set and the way negative examples are obtained) was performed over development sets sampled from the training sets. Based on this tuning, some parameters varied between the datasets and between classifier types (C_h vs. C_r). For example,

[3] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[4] http://people.csail.mit.edu/jrennie/20Newsgroups/

selection of negative examples based on cohyponyms was found useful for C_r classifiers in R10, while random examples were used in the rest of the cases.

We used *SVM^{perf}* (Joachims, 2006) with a linear kernel and binary feature weighting.

For querying the corpus we used the Lucene search engine[5] in its default setting. Up to 150 positive examples were retrieved for each classifier, with 5 examples set as the required minimum. This resulted in generating 100% of the hypothesis classifiers for both datasets and 95% and 70% of the rule classifiers for R10 and 20NG, respectively.

We computed $C_h(r)$ scores based on up to 20 sampled instances. If less than 2 examples were found in the training set, we assigned an "unknown" context match probability of 0.5, since a rare LHS occurrence does not indicate anything about its meaning in the corpus. Such cases constituted 2% (R10) and 11% (20NG) of the utilized rules.

7.3 Baseline models

To provide a more meaningful comparison with prior work, we focus on the first unsupervised step in the typical Name-based TC flow, without the subsequent supervised training. Our goal is to improve the accuracy of this first step, and we therefore compare our context model's performance to two unsupervised methods used by *Barak09*.

The first baseline, denoted *Barak$_{no\text{-}cxt}$*, is the cosine similarity score between the document and category vectors where all terms are equally weighted to a score of 1.[6] This baseline shows the performance when no context model is employed.

The second baseline, denoted *Barak$_{full}$*, is a replication of the state of the art context model for Name-based TC. In this method, LSA vectors are constructed for a document by averaging the LSA vectors of its individual terms, and for a category by averaging the LSA vectors of the terms denoting its name. The categorization score of a document-category pair is set to be the product between the cosine similarity score of the LSA vectors and the score given by the above *Barak$_{no\text{-}cxt}$* method. We note that LSA-based context models performed best also in (Gliozzo et al., 2009) and (Szpektor et al., 2008).

[5] `http://lucene.apache.org`

[6] Other attempted weighting schemes, such as tf-idf, did not yield better performance.

Model	*Reuters-10*			
	Accuracy	P	R	F_1
Barak$_{no\text{-}cxt}$	73.2	63.6	77.0	69.7
Barak$_{full}$	76.3	68.0	79.2	73.2
Class.-based	**79.3**	**71.8**	**83.6**	**77.2**
Model	*20-Newsgroups*			
	Accuracy	P	R	F_1
Barak$_{no\text{-}cxt}$	63.7	44.5	74.6	55.8
Barak$_{full}$	69.4	50.1	**82.8**	62.4
Class.-based	**73.4**	**54.7**	76.4	**63.7**

Table 1: Evaluation results.

All models were constructed based on the TC training sets, using no external corpora. The vocabulary consists of terms that appear more than once in the training set. The terms we consider include nouns, verbs, adjectives and adverbs, as well as nominal multi-word expressions.

8 Results and Analysis

Given a document, all categories for which a lexical match was found in the document are considered, and the document is classified to the highest scoring category. If all categories are assigned non-positive scores, the document is not assigned to any of them.

Based on this requirement that a document contains at least one match for the category, 4862 document-category pairs were considered for classification in R10 and 9955 pairs in 20NG. We evaluated our context model, as well as the baselines, based on the *accuracy* of these classifications, i.e. the percentage of correct decisions among the candidate document-category pairs. We also measured the models' performance in terms of micro-averaged *precision* (P), *relative recall* (R) and F_1. Like *Barak09*, recall is computed relative to the potential recall of the rule-set which provides the entailing terms.

Table 1 presents the evaluation results. As in *Barak09*, the LSA-based model outperforms the first baseline, supporting its usefulness as a context model. In both datasets our model outperformed the baselines in terms of accuracy. This result is statistically significant with $p < 0.01$ according to McNemar's test (McNemar, 1947). Recall is lower for our model in 20NG but F_1 scores are higher for both datasets. These results indicate that the classification-based context model provides a favorable alternative to the

Removed	*Reuters-10*		*20-Newsgroups*	
	Accuracy	F_1	Accuracy	F_1
-	79.3	77.2	73.4	63.7
$C_h(t)$	76.2	72.3	71.9	61.0
$C_r(t)$	80.5	77.6	74.3	64.5
$C_h(r)$	78.4	75.7	73.1	63.4

Table 2: Ablation tests results.

state of the art LSA-based method.

Table 2 presents ablation tests of our model. In each test we measured the classification performance when one of the three classification scores is ignored. Clearly, $C_h(t)$ is the most beneficial component, and in general the category classifiers help improving overall performance. The limited performance of C_r may be related to higher ambiguity in rules relative to category names, resulting in noisier training data. In addition, the small size of the training set limits the number of training examples for rule classifiers. This problem affects C_r more than C_h since, by nature, the corpus includes more occurrences of category names. Still, C_r contributes to improved recall (this fact is not visible in Table 2).

The coverage of the utilized rule-set determines the maximal (absolute) recall that can be achieved by any model. With the rule-set we used in this experiment, the recall upper bound was 59.1% for R10 and 40.6% for 20NG. However, rule coverage affects precision as well: In many cases documents are assigned to incorrect categories because the correct category is not even a candidate as no entailing term was matched for it in the document. For instance, a document with the sentence "*For sale or trade!!! BMW R60US...*" was classified by our method to the category forsale, while its gold-standard category is motorcycles. Yet, none of the rules in our rule-set triggered motorcycles as a candidate category for this document. Ideally, a context model would rule out all incorrect candidate categories; in practice even a single low score for one of the competing categories results in a false positive error in such cases (in addition to the recall loss). To reduce these problems we intend to employ additional knowledge resources in future work.

Our algorithm for retrieving training examples turned out to be not sufficiently accurate, particularly for negative examples. This is a challenging task that requires further research. Although useful for some classifier types, the use of cohyponyms may retrieve potentially positive examples as negative ones, since terms that are considered cohyponyms in WordNet are often perceived as near synonyms in common usage, e.g. *buyout* and *purchase* in the context of acquisitions. Likewise, using WordNet senses to determine ambiguity is also inaccurate. Rare or too fine-grained senses, common in WordNet, cause a term to be considered ambiguous, which in turn triggers the use of less accurate retrieval methods. For example, *auction* has a bridge-related WordNet sense which is irrelevant for our dataset, but made the term be considered ambiguous. This calls for development of other methods for determining word ambiguity, which consider the actual usage of terms in the domain rather than relying solely on WordNet.

9 Conclusions

In this paper we presented a generic classification-based scheme for comprehensively addressing context matching in textual inference scenarios. We presented a concrete implementation of the proposed scheme for Name-based TC, and showed how CP decisions can be integrated within the TC setting.

Utilizing classifiers for context matching offers several advantages. They naturally incorporate directionality and allow integrating various types of information, including ones not used in this work such as syntactic features. Our results indeed support this approach. Still, further research is required regarding issues raised by the use of multiple classifiers, scalability in particular.

Hypotheses in TC are available in advance. While also the case in other applications, it constitutes a practical challenge when hypotheses are given "online", like Information Retrieval queries, since classifiers will have to be generated on the fly. We intend to address this issue in future work.

Lastly, we plan to apply the generic classification-based approach to address context matching in other inference-based applications.

Acknowledgments

This work was partially supported by the Israel Science Foundation grant 1112/08 and the NEGEV project (`www.negev-initiative.org`).

References

Libby Barak, Ido Dagan, and Eyal Shnarch. 2009. Text Categorization from Category Name via Lexical Reference. In *HLT-NAACL (Short Papers)*.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative Learning of Selectional Preference from Unlabeled Text. In *In Proceedings of EMNLP*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Michael Connor and Dan Roth. 2007. Context Sensitive Paraphrasing with a Global Unsupervised Classifier. In *Proceedings of ECML*.

Ido Dagan, Oren Glickman, Alfio Gliozzo, Efrat Marmorshtein, and Carlo Strapparava. 2006. Direct Word Sense Matching for Lexical Substitution. In *Proceedings of COLING-ACL*.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing Textual Entailment: Rational, Evaluation and Approaches. *Natural Language Engineering*, pages 15(4):1–17.

Scott Deerwester, Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391–407.

Georgiana Dinu and Mirella Lapata. 2010. Topic Models for Meaning Similarity in Context. In *Proceedings of Coling 2010: Posters*.

Doug Downey and Oren Etzioni. 2009. Look Ma, No Hands: Analyzing the Monotonic Feature Abstraction for Text Classification. In *Proceedings of NIPS*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One Sense per Discourse. In *Proceedings of the workshop on Speech and Natural Language*.

Alfio Gliozzo, Carlo Strapparava, and Ido Dagan. 2009. Improving Text Categorization Bootstrapping via Unsupervised Learning. *ACM Trans. Speech Lang. Process.*, 6:1:1–1:24, October.

Sanda M. Harabagiu, Steven J. Maiorano, and Marius A. Paşca. 2003. Open-domain Textual Question Answering Techniques. *Natural Language Engineering*, 9:231–267, September.

Thorsten Joachims. 2006. Training Linear SVMs in Linear Time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of HLT-NAACL*.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional Distributional Similarity for Lexical Inference. *Natural Language Engineering*, 16(4):359–389.

Diana McCarthy and Roberto Navigli. 2009. The English Lexical Substitution Task. *Language Resources and Evaluation*, 43(2):139–159.

Quinn McNemar. 1947. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, June.

Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning Inferential Selectional Preferences. In *Proceedings of NAACL-HLT*.

Marco Pennacchiotti, Roberto Basili, Diego De Cao, and Paolo Marocco. 2007. Learning Selectional Preferences for Entailment or Paraphrasing Rules. In *Proceedings of RANLP*.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A Latent Dirichlet Allocation Method for Selectional Preferences. In *Proceedings of ACL*.

Eyal Shnarch, Libby Barak, and Ido Dagan. 2009. Extracting Lexical Reference Rules from Wikipedia. In *Proceedings of IJCNLP-ACL*.

Noah A. Smith and Jason Eisner. 2005. Contrastive Estimation: Training Log-linear Models on Unlabeled Data. In *Proceedings of ACL*.

Idan Szpektor, Ido Dagan, Roy Bar-Haim, and Jacob Goldberger. 2008. Contextual Preferences. In *Proceedings of ACL-08: HLT*.

Chapter 6

Assessing the Role of Discourse References in Entailment Inference

Assessing the Role of Discourse References in Entailment Inference

Shachar Mirkin, Ido Dagan
Bar-Ilan University
Ramat-Gan, Israel
{mirkins,dagan}@cs.biu.ac.il

Sebastian Padó
University of Stuttgart
Stuttgart, Germany
pado@ims.uni-stuttgart.de

Abstract

Discourse references, notably coreference and bridging, play an important role in many text understanding applications, but their impact on textual entailment is yet to be systematically understood. On the basis of an in-depth analysis of entailment instances, we argue that discourse references have the potential of substantially improving textual entailment recognition, and identify a number of research directions towards this goal.

1 Introduction

The detection and resolution of *discourse references* such as coreference and bridging anaphora play an important role in text understanding applications, like question answering and information extraction. There, reference resolution is used for the purpose of combining knowledge from multiple sentences. Such knowledge is also important for Textual Entailment (TE), a generic framework for modeling semantic inference. TE reduces the inference requirements of many text understanding applications to the problem of determining whether the meaning of a given textual assertion, termed *hypothesis* (H), can be inferred from the meaning of certain *text* (T) (Dagan et al., 2006).

Consider the following example:

(1) T: "*Not only had he developed an aversion to the* ***President***$_1$ *and politics in general,* ***Oswald***$_2$ *was also a failure with Marina, his wife. [...] Their relationship was supposedly responsible for why* ***he***$_2$ *killed* ***Kennedy***$_1$."

H: "*Oswald killed President Kennedy.*"

The understanding that the second sentence of the text entails the hypothesis draws on two coreference relationships, namely that *he* is *Oswald*, and that the *Kennedy* in question is *President Kennedy*. However, the utilization of discourse information for such inferences has been so far limited mainly to the substitution of nominal coreferents, while many aspects of the interface between discourse and semantic inference needs remain unexplored.

The recently held Fifth Recognizing Textual Entailment (RTE-5) challenge (Bentivogli et al., 2009a) has introduced a *Search* task, where the text sentences are interpreted in the context of their full discourse, as in Example 1 above. Accordingly, TE constitutes an interesting framework – and the Search task an adequate dataset – to study the interrelation between discourse and inference.

The goal of this study is to analyze the roles of discourse references for textual entailment inference, to provide relevant findings and insights to developers of both reference resolvers and entailment systems and to highlight promising directions for the better incorporation of discourse phenomena into inference. Our focus is on a manual, in-depth assessment that results in a classification and quantification of discourse reference phenomena and their utilization for inference. On this basis, we develop an account of formal devices for incorporating discourse references into the inference computation. An additional point of interest is the interrelation between entailment knowledge and coreference. E.g., in Example 1 above, knowing that Kennedy was a president can alleviate the need for coreference resolution. Conversely, coreference resolution can often be used to overcome gaps in entailment knowledge.

Structure of the paper. In Section 2, we provide background on the use of discourse references in natural language processing (NLP) in general and specifically in TE. Section 3 describes the goals of this study, followed by our analysis scheme (Section 4) and the required inference

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1209–1219,
Uppsala, Sweden, 11-16 July 2010.

mechanisms (Section 5). Section 6 presents quantitative findings and further observations. Conclusions are discussed in Section 7.

2 Background

2.1 Discourse in NLP

Discourse information plays a role in a range of NLP tasks. It is obviously central to *discourse processing* tasks such as text segmentation (Hearst, 1997). Reference information provided by discourse is also useful for *text understanding* tasks such as question answering (QA), information extraction (IE) and information retrieval (IR) (Vicedo and Ferrndez, 2006; Zelenko et al., 2004; Na and Ng, 2009), as well as for the acquisition of lexical-semantic "narrative schema" knowledge (Chambers and Jurafsky, 2009). Discourse references have been the subject of attention in both the Message Understanding Conference (Grishman and Sundheim, 1996) and the Automatic Content Extraction program (Strassel et al., 2008).

The simplest form of information that discourse provides is *coreference*, i.e., information that two linguistic expressions refer to the same entity or event. Coreference is particularly important for processing pronouns and other anaphoric expressions, such as *he* in Example 1. Ability to resolve this reference translates directly into, e.g., a QA system's ability to answer questions like *Who killed Kennedy?*.

A second, more complex type of information stems from *bridging references*, such as in the following discourse (Asher and Lascarides, 1998):

(2) "*I've just **arrived**. The **camel** is outside.*"

While coreference indicates equivalence, bridging points to the existence of a salient semantic relation between two distinct entities or events. Here, it is (informally) *'means of transport'*, which would make the discourse (2) relevant for a question like *How did I arrive here?*. Other types of bridging relations include set-membership, roles in events and consequence (Clark, 1975).

Note, however, that text understanding systems are generally limited to the resolution of entity (or even just pronoun) coreference, e.g. (Li et al., 2009; Dali et al., 2009). An important reason is the unavailability of tools to resolve the more complex (and difficult) forms of discourse reference such as event coreference and bridging.[1] Another reason is uncertainty about their practical importance.

2.2 Discourse in Textual Entailment

Textual Entailment has been introduced in Section 1 as a common-sense notion of inference. It has spawned interest in the computational linguistics community as a common denominator of many NLP tasks including IE, summarization and tutoring (Romano et al., 2006; Harabagiu et al., 2007; Nielsen et al., 2009).

Architectures for Textual Entailment. Over the course of recent RTE challenges (Giampiccolo et al., 2007; Giampiccolo et al., 2008), the main benchmark for TE technology, two architectures for modeling TE have emerged as dominant: *transformations* and *alignment*. The goal of ***transformation***-based TE models is to determine the entailment relation $T \Rightarrow H$ by finding a "proof", i.e., a sequence of consequents, $(T, T_1, \ldots, T_n)$, such that T_n=H (Bar-Haim et al., 2008; Harmeling, 2009), and that in each transformation, $T_i \rightarrow T_{i+1}$, the consequent T_{i+1} is entailed by T_i. These transformations commonly include lexical modifications and the generation of syntactic alternatives. The second major approach constructs an ***alignment*** between the linguistic entities of the trees (or graphs) of T and H, which can represent syntactic structure, semantic structure, or non-hierarchical phrases (Zanzotto et al., 2009; Burchardt et al., 2009; MacCartney et al., 2008). H is assumed to be entailed by T if its entities are aligned "well" to corresponding entities in T. Alignment quality is generally determined based on features that assess the validity of the local replacement of the T entity by the H entity.

While transformation- and alignment-based entailment models look different at first glance, they ultimately have the same goal, namely obtaining a maximal *coverage* of H by T, i.e. to identify matches of as many elements of H within T as possible.[2] To do so, both architectures typically make use of *inference rules* such as '*Y was purchased by X* $\rightarrow$ *X paid for Y*', either by directly applying them as transformations, or by using them

[1] Some studies, e.g. (Markert et al., 2003; Poesio et al., 2004), address the resolution of a few specific kinds of bridging relations; yet, wide-scope systems for bridging resolution are unavailable.

[2] Clearly, the details of how the final entailment decision is made based on the attained coverage differ substantially among models.

to score alignments. Rules are generally drawn from external knowledge resources, such as WordNet (Fellbaum, 1998) or DIRT (Lin and Pantel, 2001), although knowledge gaps remain a key obstacle (Bos, 2005; Balahur et al., 2008; Bar-Haim et al., 2008).

Discourse in previous RTE challenges. The first two rounds of the RTE challenge used "self-contained" texts and hypotheses, where discourse considerations played virtually no role. A first step towards a more comprehensive notion of entailment was taken with RTE-3 (Giampiccolo et al., 2007), when paragraph-length texts were first included and constituted 17% of the texts in the test set. Chambers et al. (2007) report that in a sample of $T - H$ pairs drawn from the development set, 25% involved discourse references.

Using the concepts introduced above, the impact of discourse references can be generally described as a *coverage problem*, independent of the system's architecture. In Example 1, the hypothesis word *Oswald* cannot be safely linked to the text pronoun *he* without further knowledge about *he*; the same is true for '*Kennedy* → *President Kennedy*' which involves a specialization that is only warranted in the specific discourse.

A number of systems have tried to address the question of coreference in RTE as a preprocessing step prior to inference proper, with most systems using off-the-shelf coreference resolvers such as JavaRap (Qiu et al., 2004) or OpenNLP[3]. Generally, anaphoric expressions were textually replaced by their antecedents. Results were inconclusive, however, with several reports about errors introduced by automatic coreference resolution (Agichtein et al., 2008; Adams et al., 2007). Specific evaluations of the contribution of coreference resolution yielded both small negative (Bar-Haim et al., 2008) and insignificant positive (Chambers et al., 2007) results.

3 Motivation and Goals

The results of recent studies, as reported in Section 2.2, seem to show that current resolution of discourse references in RTE systems hardly affects performance. However, our intuition is that these results can be attributed to four major limitations shared by these studies: (1) the datasets, where discourse phenomena were not well represented; (2) the off-the-shelf coreference resolution systems which may have been not robust enough; (3) the limitation to nominal coreference; and (4) overly simple integration of reference information into the inference engines.

The goal of this paper is to assess the impact of discourse references on entailment with an annotation study which removes these limitations. To counteract (1), we use the recent RTE-5 Search dataset (details below). To avoid (2), we perform a manual analysis, assuming discourse references as predicted by an oracle. With regards to (3), our annotation scheme covers coreference and bridging relations of all syntactic categories and classifies them. As for (4), we suggest several operations necessary to integrate the discourse information into an entailment engine.

In contrast to the numerous existing datasets annotated for discourse references (Hovy et al., 2006; Strassel et al., 2008), we do not annotate exhaustively. Rather, we are interested specifically in those references instances that impact inference. Furthermore, we analyze each instance from an entailment perspective, characterizing the relevant factors that have an impact on inference. To our knowledge, this is the first such in-depth study.[4]

The results of our study are of twofold interest. First, they provide guidance for the developers of reference resolvers who might prioritize the scope of their systems to make them more valuable for inference. Second, they point out potential directions for the developers of inference systems by specifying what additional inference mechanisms are needed to utilize discourse information.

The RTE-5 Search dataset. We base our annotation on the Search task dataset, a new addition to the recent Fifth RTE challenge (Bentivogli et al., 2009a) that is motivated by the needs of NLP applications and drawn from the TAC summarization track. In the Search task, TE systems are required to find *all* individual sentences in a given corpus which entail the hypothesis – a setting that is sensible not only for summarization, but also for information access tasks like QA. Sentences are judged individually, but "are to be interpreted in the context of the corpus as they rely on explicit and implicit references to entities, events, dates, places, etc., mentioned elsewhere in the corpus" (Bentivogli et al., 2009b).

[3] http://opennlp.sourceforge.net

[4] The guidelines and the dataset are available at http://www.cs.biu.ac.il/~nlp/downloads/

		Text	Hypothesis
i	T'	*Once the reform becomes law, Spain will join the Netherlands and Belgium in allowing* ***homosexual marriages.***	*Massachusetts allows* ***homosexual marriages***
	T	***Such unions*** *are also legal in six Canadian provinces and the northeastern US state of Massachusetts.*	
ii	T'	*The official name of* ***2003 UB313*** *has yet to be determined.*	***2003 UB313*** *is in the Kuiper Belt*
	T	*Brown said he expected to find a moon orbiting* ***Xena*** *because many Kuiper Belt objects are paired with moons.*	
iii	T'_a	*All seven aboard the* ***AS-28 submarine*** *appeared to be in satisfactory condition, naval spokesman said.*	*The* ***AS-28 mini submarine*** *was trapped underwater*
	T'_b	*British crews were working with Russian naval authorities to maneuver the unmanned robotic vehicle and untangle the* ***AS-28****.*	
	T	*The Russian military was racing against time early Friday to rescue a* ***mini submarine trapped on the seabed.***	
iv	T'	***China*** *seeks solutions to its coal mine safety.*	*A* ***mining accident in China*** *has killed several miners*
	T	*A* ***recent accident*** *has cost more than a dozen miners their lives.*	
v	T''	*A remote-controlled device was lowered to the stricken vessel to cut the cables in which the* ***AS-28 vehicle*** *is caught.*	*The* ***AS-28 mini submarine*** *was trapped underwater*
	T'	*The* ***mini submarine*** *was resting on the seabed at a depth of about 200 meters.*	
	T	*Specialists said* ***it*** *could have become tangled up with a metal cable or in sunken nets from a fishing trawler.*	
vi	T	*. . . dried up* ***lakes in Siberia****, because the permafrost beneath* ***them*** *has begun to thaw.*	*The ice is* ***melting in the Arctic***

Table 1: Examples for discourse-dependent entailment in the RTE-5 dataset, where the inference of H depends on reference information from the discourse sentences T' / T''. Referring terms (in T) and target terms (in H) are shown in boldface.

4 Analysis Scheme

For annotating the RTE-5 data, we operationalize reference relations that are *relevant* for entailment as those that improve *coverage*. Recall from Section 2.2 that the concept of coverage is applicable to both transformation and alignment models, all of which aim at maximizing coverage of H by T.

We represent T and H as syntactic trees, as common in the RTE literature (Zanzotto et al., 2009; Agichtein et al., 2008). Specifically, we assume MINIPAR-style (Lin, 1993) dependency trees where nodes represent text expressions and edges represent the syntactic relations between them. We use "term" to refer to text expressions, and "components" to refer to nodes, edges, and subtrees. Dependency trees are a popular choice in RTE since they offer a fairly semantics-oriented account of the sentence structure that can still be constructed robustly. In an ideal case of entailment, all nodes *and* dependency edges of H are covered by T.

For each $T - H$ pair, we annotate all relevant discourse references in terms of three items: the *target component* in H, the *focus term* in T, and the *reference term* which stands in a reference relation to the focus term. By resolving this reference, the target component can usually be inferred; sometimes, however, more than one reference term needs to be found. We now define and illustrate these concepts on examples from Table 1.[5]

The ***target component*** is a tree component in H that cannot be covered by the "local" material from T. An example for a tree component is Example (v), where the target component *AS-28 mini submarine* in H cannot be inferred from the pronoun *it* in T. Example (vi) demonstrates an edge as target component. In this case, the edge in H connecting *melt* with the modifier *in the Arctic* is not found in T. Although each of the hypothesis' nodes can be covered separately via knowledge-based rules (e.g. '*Siberia* → *Arctic*', '*permafrost* → *ice*', '*thaw* ↔ *melt*'), the resulting fragments in T are unconnected without the (intra-sentential) coreference between *them* and *lakes in Siberia*.

For each target component, we identify its ***focus term*** as the expression in T that does not cover the target component itself but participates in a reference relation that can help covering it.

We follow the focus term's reference chain to a ***reference term*** which can, either separately or in combination with the focus term, help covering the target component. In Example (ii), where the

[5] In our annotation, we assume throughout that some knowledge about basic admissible transformations is available, such as passive to active or derivational transformations; for brevity, we ignore articles in the examples and treat named entities as single nodes.

target component in H is *2003 UB313*, *Xena* is the focus term in T and the reference term is a mention of *2003 UB313* in a previous sentence, T'. In this case, the reference term covers the entire target component on its own.

An additional attribute that we record for each instance is whether resolving the discourse reference is *mandatory* for determining entailment, or *optional*. In Example (v), it is mandatory: the inference cannot be completed without the knowledge provided by the discourse. In contrast, in Example (ii), inferring *2003 UB313* from *Xena* is optional. It can be done either by identifying their coreference relation, or by using background knowledge in the form of an entailment rule, '*Xena* $\leftrightarrow$ *2003 UB313*', that is applicable in the context of astronomy. Optional discourse references represent instances where discourse information and TE knowledge are interchangeable. As mentioned, knowledge gaps constitute a major obstacle for TE systems, and we cannot rely on the availability of any ceratin piece of knowledge to the inference process. Thus, in our scheme, mandatory references provide a "lower bound" with regards to the necessity to resolve discourse references, even in the presence of *complete knowledge*; optional references, on the other hand, set an "upper bound" for the contribution of discourse resolution to inference, when *no knowledge* is available. At the same time, this scheme allows investigating how much TE knowledge can be replaced by (perfect) discourse processing.

When choosing a reference term, we search the reference chain of the focus term for the nearest expression that is identical to the target component or a subcomponent of it. If we find such an expression, covering the identical part of the target component requires no entailment knowledge. If no identical reference term exists, we choose the semantically 'closest' term from the reference chain, i.e. the term which requires the least knowledge to infer the target component. For instance, we may pick *permafrost* as the semantically closet term to the target *ice* if the latter is not found in the focus term's reference chain.

Finally, for each reference relation that we annotate, we record four additional attributes which we assumed to be informative in an evaluation. First, the *reference type*: Is the relation a coreference or a bridging reference? Second, the *syntactic type* of the focus and reference terms. Third, the *focus/reference terms entailment status* – does some kind of entailment relation hold between the two terms? Fourth, the *operation* that should be performed on the focus and reference terms to obtain coverage of the target component (as specified in Section 5).

5 Integrating Discourse References into Entailment Recognition

In initial analysis we found that the standard substitution operation applied by virtually all previous studies for integrating coreference into entailment is insufficient. We identified three distinct cases for the integration of discourse reference knowledge in entailment, which correspond to different relations between the target component, the focus term and the reference term. This section describes the three cases and characterizes them in terms of tree transformations. An initial version of these transformations is described in (Abad et al., 2010). We assume a transformation-based entailment architecture (cf. Section 2.2), although we believe that the key points of our account are also applicable to alignment-based architecture. Transformations create revised trees that cover previously uncovered target components in H. The output of each transformation, T_1, is comprised of copies of the components used to construct it, and is appended to the discourse forest, which includes the dependency trees of all sentences and their generated consequents.

We assume that we have access to a dependency tree for H, a dependency forest for T and its discourse context, as well as the output of a perfect discourse processor, i.e., a complete set of both coreference and bridging relations, including the type of bridging relation (e.g. *part-of*, *cause*).

We use the following notation. We use x, y for tree nodes, and S_x to denote a (sub-)tree with root x. $lab(x)$ is the label of the incoming edge of x (i.e., its grammatical function). We write $C(x, y)$ for a coreference relation between S_x and S_y, the corresponding trees of the focus and reference terms, respectively. We write $B_r(x, y)$ for a bridging relation, where r is its type.

(1) Substitution: This is the most intuitive and widely-used transformation, corresponding to the treatment of discourse information in existing systems. It applies to coreference relations, when an expression found elsewhere in the text (the reference term) can cover all missing information (the

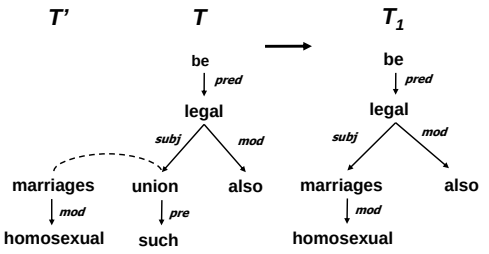

Figure 1: The *Substitution* transformation, demonstrated on the relevant subtrees of Example (i). The dashed line denotes a discourse reference.

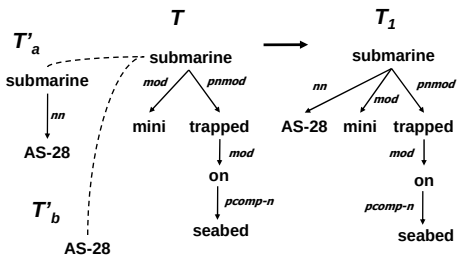

Figure 2: The *dependent-merge* (T'_a) and *head-merge* (T'_b) transformations (Example (iii)).

target component) on its own. In such cases, the reference term can replace the entire focus term. Apparently (cf. Section 6), substitution applies also to some types of bridging relations, such as *set-membership*, when the member is sufficient for representing the entire set for the necessary inference. For example, in "*I met two people yesterday. The woman told me a story.*" (Clark, 1975), substituting *two people* with *woman* results in a text which is entailed from the discourse, and which allows inferring "*I met a woman yesterday.*"

In a parse tree representation, given a coreference relation $C(x, y)$ (or $B_r(x, y)$), the newly generated tree, T_1, consists of a copy of T, where the entire tree S_x is replaced by a copy of S_y . In Figure 1, which shows Example (i) from Table 1, *such unions* is substituted by *homosexual marriages*.

Head-substitution. Occasionally, substituting only the head of the focus term is sufficient. In such cases, only the root nodes x and y are substituted. This is the case, for example, with synonymous verbs with identical subcategorization frames (like *melt* and *thaw*). As verbs typically constitute tree roots in dependency parses, substituting or merging (see below) their entire trees might be inappropriate or wasteful. In such cases, the simpler head-substitution may be applied.

(2) Merge: In contrast to substitution, where a match for the entire target component is found elsewhere in the text, this transformation is required when parts of the missing information are scattered among multiple locations in the text. We distinguish between two types of merge transformations: (a) *dependent-merge*, and (b) *head-merge*, depending on the syntactic roles of the merged components.

(a) Dependent-Merge. This operation is applicable when the head of either the focus or reference terms (of both) matches the head node of the target component, but modifiers from both of them are required to cover the target component's dependents. The modifiers are therefore merged as dependents of a single head node, to create a tree that covers the entire target component. Dependent-merge is illustrated in Figure 2, using Example (iii). The component we wish to cover in H is the noun phrase *AS-28 mini submarine*. Unfortunately, the focus term in T, "*mini submarine trapped on the seabed*", covers only the modifier *mini*, but not *AS-28*. This modifier can however be provided by the coreferent term in T'_a (left upper corner). Once merged, the inference engine can, e.g., employ the rule '*on seabed* $\rightarrow$ *underwater*' to cover H completely.

Formally, assume without loss of generality that y, the reference term's head, matches the root node of the target component. Given $C(x, y)$, we define T_1 as a copy of T, where (i) the subtree S_x is replaced by S_y, and (ii) for all children c of x, a copy of S_c is placed under the copy of y in T_1 with its original edge label, $lab(c)$.

(b) Head-merge. An alternative way to recover the missing information in Example (iii) is to find a reference term whose head word itself (rather than one of its modifiers) matches the target component's missing dependent, as with *AS-28* in Figure 2 in the bottom left corner (T'_b). In terms of parse trees, we need to add one tree as a dependent of the other. Formally, given $C(x, y)$, similarly to dependent-merge, T_1 is created as a copy of T where the subtree S_x is replaced by either S_x or S_y, depending on whichever of x and y matches the target component's head. Assume it is x, for example. Then, a copy of S_y is added as a new child to x. In our sample, head-merge operations correspond to internal coreferences within nominal target components (such as between *AS-28* and *mini submarine* in this case). The appropriate label, $lab(y)$, in these cases is nn (nominal modi-

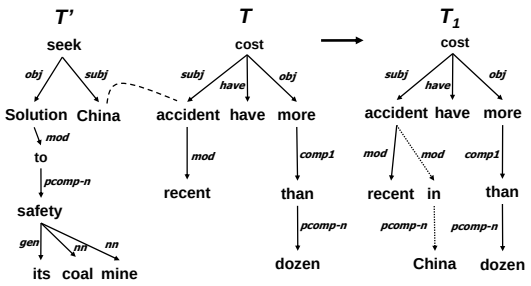

Figure 3: The *insertion* transformation. Dotted edges mark the newly inserted path (Ex. (iv)).

fier). Further analysis is required to specify what other dependencies can hold between such coreferring heads.

(3) Insertion: The last transformation, insertion, is used when a relation that is realized in H is missing from T and is only implied via a bridging relation. In Example (iv), the location that is explicitly mentioned in H can only be covered by T by resolving a bridging reference with *China* in T'. To connect the bridging referents, a new tree component representing the bridging relation is inserted into the consequent tree T_1. In this example, the component connects *China* and *recent accident* via the *in* preposition. Formally, given a bridging relation $B_r(x, y)$, we introduce a new subtree S^r_z into T_1, where z is a child of x and $lab(z) = lab_r$. S^r_z must contain a variable node that is instantiated with a copy of $S(y)$.

This transformation stands out from the others in that it introduces new material. For each bridging relation, it adds a specific subtrees S^r via an edge labeled with lab_r. These two items form the dependency representation of the bridging relation B_r and must be provided by the interface between the discourse and the inference systems. Clearly, their exact form depends on the set of bridging relations provided by the discourse resolver as well as the details of the dependency parses.

As shown in Figure 3, the bridging relation *located-in* (r) is represented by inserting a subtree S^r_z headed by *in* (z) into T_1 and connecting it to *accident* (x) as a modifier (lab_r). The subtree S^r_z consists of a variable node which is connected to *in* with a *pcomp-n* dependency (a nominal head of a prepositional phrase), and which is instantiated with the node *China* (y) when the transformation is applied. Note that the structure of S^r_z and the way it is inserted into T_1 are predefined by the abovementioned interface; only the node to which it is attached and the contents of the variable node are determined at transformation-time.

As another example, consider the following short text from (Clark, 1975): *John was murdered yesterday. The knife lay nearby*. Here, the bridging relation between the murder event and the instrument, *the knife* (x), can be addressed by inserting under x a subtree for the clause *with which* as S^r_z, with a variable which is instantiated by the parse-tree (headed by *murdered*, y) of the entire first sentence *John was murdered yesterday*.

Transformation chaining. Since our transformations are defined to be minimal, some cases require the application of multiple transformations to achieve coverage. Consider Example (v), Table 1. We wish to cover *AS-28 mini submarine* in H from the coreferring *it* in T, *mini submarine* in T' and *AS-28 vehicle* in T''. A substitution of *it* by either coreference does not suffice, since none of the antecedents contains all necessary modifiers. It is therefore necessary to substitute *it* first by one of the coreferences and then merge it with the other.

6 Results

We analyzed 120 sentence-hypothesis pairs of the RTE-5 development set (21 different hypotheses, 111 distinct sentences, 53 different documents). Below, we summarize our findings, focusing on the relation between our findings and the assumptions of previous studies as discussed in Section 3.

General statistics. We found that 44% of the pairs contained reference relations whose resolution was mandatory for inference. In another 28%, references could optionally support the inference of the hypothesis. In the remaining 28%, references did not contribute towards inference. The total number of relevant references was 137, and 37 pairs (27%) contained multiple relevant references. These numbers support our assumption that discourse references play an important role in inference.

Reference types. 73% of the identified references are coreferences and 27% are bridging relations. The most common bridging relation was the location of events (e.g. *Arctic* in ice melting events), generally assumed to be known throughout the document. Other bridging relations we encountered include cause (e.g. between *injured* and *attack*), event participants and set membership.

(%)	Pronoun	NE	NP	VP
Focus term	9	19	49	23
Reference term	-	43	43	14

Table 2: Syntactic types of discourse references

(%)	Sub.	Merge	Insertion
Coreference	62	38	-
Bridging	30	-	70
Total	54	28	18

Table 3: Distribution of transformation types

Syntactic types. Table 2 shows that 77% of all focus terms and 86% of the reference terms were nominal phrases, which justifies their prominent position in work on anaphora and coreference resolution. However, almost a quarter of the focus terms were verbal phrases. We found these focus terms to be frequently crucial for entailment since they included the main predicate of the hypothesis.[6] This calls for an increased focus on the resolution of event references.

Transformations. Table 3 shows the relative frequencies of all transformations. Again, we found that the "default" transformation, substitution, is the most frequent one, and is helpful for both coreference and bridging relations. Substitution is particularly useful for handling pronouns (14% of all substitution instances), the replacement of named entities by synonymous names (32%), the replacement of other NPs (38%), and the substitution of verbal head nodes in event coreference (16%). Yet, in nearly half the cases, a different transformation had to be applied. Insertion accounts for the majority of bridging cases. Head-merge is necessary to integrate proper nouns as modifiers of other head nouns. Dependent-merge, responsible for 85% of the merge transformations, can be used to complete nominal focus terms with missing modifiers (e.g., adjectives), as well as for merging other dependencies between coreferring predicates. This result indicates the importance of incorporating other transformations into inference systems.

Distance of reference terms. The distance between the focus and the reference terms varied considerably, ranging from intra-sentential reference relations and up to several dozen sentences. For more than a quarter of the focus terms, we had to go to other documents to find reference terms that, possibly in conjunction with the focus term, could cover the target components. Interestingly, all such cases involved coreference (about equally divided between the merge transformations and substitutions), while bridging was always "document-local". This result reaffirms the usefulness of cross-document coreference resolution for inference (Huang et al., 2009).

Discourse resolution as preprocessing? In existing RTE systems, discourse references are typically resolved as a preprocessing step. While our annotation was manual and cannot yield direct results about processing considerations, we observed that discourse relations often hold between complex, and deeply embedded, expressions, which makes their automatic resolution difficult. Of course, many RTE systems attempt to normalize and simplify H and T, e.g., by splitting conjunctions or removing irrelevant clauses, but these operations are usually considered a part of the inference rather the preprocessing phase (cf. e.g., Bar-Haim et al. (2007)). Since the resolution of discourse references is likely to profit from these steps, it seems desirable to "postpone" it until after simplification. In transformation-based systems, it might be natural to add discourse-based transformations to the set of inference operations, while in alignment-based systems, discourse references can be integrated into the computation of alignment scores.

Discourse references vs. entailment knowledge. We have stated before that even if a discourse reference is not strictly necessary for entailment, it may be interesting because it represents an alternative to the use of knowledge rules to cover the hypothesis. Sometimes, these rules are generally applicable (e.g., '*Alaska* → *Arctic*'). However, often they are context-specific. Consider the following sentence as T for the hypothesis H: "*The ice is melting in the Arctic*":

(3) T: "*The scene at the* ***receding*** *edge of the Exit Glacier was part festive gathering, part nature tour with an apocalyptic edge.*"

While it is possible to cover *melting* using a rule '*melting* ↔ *receding*', this rule is only valid under quite specific conditions (e.g., for the subject *ice*). Instead of determining the applicability of the rule, a discourse-aware system can take the next sen-

[6] The lower proportion of VPs among reference terms stems from bridging relations between VPs and nominal dependents, such as the abovementioned "location" relation.

tence into account, which contains a coreferring event to *receding* that can cover *melting* in H:

(4) T': "*... people moved closer to the rope line near the glacier as it shied away, practically groaning and* ***melting*** *before their eyes.*"

Discourse relations can in fact encode arbitrarily complex world knowledge, as in the following pair:

(5) H: "*The* ***serial*** *killer BTK was accused of at least 7 killings starting in the 1970's.*"

T: "*Police say BTK may have killed as many as 10 people between 1974 and 1991.*"

Here, the H modifier *serial*, which does not occur in T, can be covered either by world knowledge (a person who killed 10 people is a serial killer), or by resolving the coreference of *BTK* to the term *the serial killer BTK* which occurs in the discourse around T. Our conclusion is that not only can discourse references often replace world knowledge in principle, in practice it often seems easier to resolve discourse references than to determine whether a rule is applicable in a given context or to formalize complex world knowledge as inference rules. Our annotation provides further empirical support to this claim: An entailment relation exists between the focus and reference terms in 60% of the focus-reference term pairs, and in many of the remainder, entailment holds between the terms' heads. Thus, discourse provides relations which are many times equivalent to entailment knowledge rules and can therefore be utilized in their stead.

7 Conclusions

This work has presented an analysis of the relation between discourse references and textual entailment. We have identified a set of limitations common to the handling of discourse relations in virtually all entailment systems. They include the use of off-the-shelf resolvers that concentrate on nominal coreference, the integration of reference information through substitution, and the RTE evaluation schemes, which played down the role of discourse. Since in practical settings, discourse plays an important role, our goal was to develop an agenda for improving the handling of discourse references in entailment-based inference.

Our manual analysis of the RTE-5 dataset shows that while the majority of discourse references that affect inference are nominal coreference relations, another substantial part is made up by verbal terms and bridging relations. Furthermore, we have demonstrated that substitution alone is insufficient to extract all relevant information from the wide range of discourse references that are frequently relevant for inference. We identified three general cases, and suggested matching operations to obtain the relevant inferences, formulated as tree transformations. Furthermore, our evidence suggests that for practical reasons, the resolution of discourse references should be tightly integrated into entailment systems instead of treating it as a preprocessing step.

A particularly interesting result concerns the interplay between discourse references and entailment knowledge. While semantic knowledge (e.g., from WordNet or Wikipedia) has been used beneficially for coreference resolution (Soon et al., 2001; Ponzetto and Strube, 2006), reference resolution has, to our knowledge, not yet been employed to validate entailment rules' applicability. Our analyses suggest that in the context of deciding textual entailment, reference resolution and entailment knowledge can be seen as complementary ways of achieving the same goal, namely enriching T with additional knowledge to allow the inference of H. Given that both of the technologies are still imperfect, we envisage the way forward as a joint strategy, where reference resolution and entailment rules mutually fill each other's gaps (cf. Example 3).

In sum, our study shows that textual entailment can profit substantially from better discourse handling. The next challenge is to translate the theoretical gain into practical benefit. Our analysis demonstrates that improvements are necessary both on the side of discourse reference resolution systems, which need to cover more types of references, as well as a better integration of discourse information in entailment systems, even for those relations which are within the scope of available resolvers.

Acknowledgements

This work was partially supported by the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886 and the Israel Science Foundation grant 1112/08.

References

Azad Abad, Luisa Bentivogli, Ido Dagan, Danilo Giampiccolo, Shachar Mirkin, Emanuele Pianta, and Asher Stern. 2010. A resource for investigating the impact of anaphora and coreference on inference. In *Proceedings of LREC*.

Rod Adams, Gabriel Nicolae, Cristina Nicolae, and Sanda Harabagiu. 2007. Textual entailment through extended lexical overlap and lexico-semantic matching. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

E. Agichtein, W. Askew, and Y. Liu. 2008. Combining lexical, syntactic, and semantic evidence for textual entailment classification. In *Proceedings of TAC*.

Nicholas Asher and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1):83–113.

Alexandra Balahur, Elena Lloret, Óscar Ferrández, Andrés Montoyo, Manuel Palomar, and Rafael Muñoz. 2008. The DLSIUAES team's participation in the TAC 2008 tracks. In *Proceedings of TAC*.

Roy Bar-Haim, Ido Dagan, Iddo Greental, and Eyal Shnarch. 2007. Semantic inference at the lexical-syntactic level. In *Proceedings of AAAI*.

Roy Bar-Haim, Jonathan Berant, Ido Dagan, Iddo Greental, Shachar Mirkin, and Eyal Shnarch amd Idan Szpektor. 2008. Efficient semantic deduction and approximate matching over compact parse forests. In *Proceedings of TAC*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009a. The fifth pascal recognizing textual entailment challenge. In *Proceedings of TAC*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2009b. Considering discourse references in textual entailment annotation. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL2009)*.

Johan Bos. 2005. Recognising textual entailment with logical inference. In *Proceedings of EMNLP*.

Aljoscha Burchardt, Marco Pennacchiotti, Stefan Thater, and Manfred Pinkal. 2009. Assessing the impact of frame semantics on textual entailment. *Journal of Natural Language Engineering*, 15(4):527–550.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL-IJCNLP*.

Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine de Marneffe, Daniel Ramage, Eric Yeh, and Christopher D. Manning. 2007. Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Herbert H. Clark. 1975. Bridging. In R. C. Schank and B. L. Nash-Webber, editors, *Theoretical issues in natural language processing*, pages 169–174. Association of Computing Machinery.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Lorand Dali, Delia Rusu, Blaz Fortuna, Dunja Mladenic, and Marko Grobelnik. 2009. Question answering based on semantic graphs. In *Proceedings of the Workshop on Semantic Search (SemSearch 2009)*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2008. The fourth pascal recognizing textual entailment challenge. In *Proceedings of TAC*.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: a brief history. In *Proceedings of the 16th conference on Computational Linguistics*.

Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. 2007. Satisfying information needs with multi-document summaries. *Information Processing & Management*, 43:1619–1642.

Stefan Harmeling. 2009. Inferring textual entailment with a probabilistically sound calculus. *Journal of Natural Language Engineering*, pages 459–477.

Marti A. Hearst. 1997. Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of HLT-NAACL*.

Jian Huang, Sarah M. Taylor, Jonathan L. Smith, Konstantinos A. Fotiadis, and C. Lee Giles. 2009. Profile based cross-document coreference using kernelized fuzzy relational clustering. In *Proceedings of ACL-IJCNLP*.

Fangtao Li, Yang Tang, Minlie Huang, and Xiaoyan Zhu. 2009. Answering opinion questions with random walks on graphs. In *Proceedings of ACL-IJCNLP*.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 4(7):343–360.

Dekang Lin. 1993. Principle-based parsing without overgeneration. In *Proceedings of ACL*.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of EMNLP*.

Katja Markert, Malvina Nissim, and Natalia N. Modjeska. 2003. Using the web for nominal anaphora resolution. In *Proceedings of EACL Workshop on the Computational Treatment of Anaphora*.

Seung-Hoon Na and Hwee Tou Ng. 2009. A 2-poisson model for probabilistic coreference of named entities for improved text retrieval. In *Proceedings of SIGIR*.

Rodney D. Nielsen, Wayne Ward, and James H. Martin. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(4):479–501.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of ACL*.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of HLT*.

Long Qiu, Min-Yen Kan, and Tat-Seng Chua. 2004. A public reference implementation of the rap anaphora resolution algorithm. In *Proceedings of LREC*.

Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. 2006. Investigating a generic paraphrase-based approach for relation extraction. In *Proceedings of EACL*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Stephanie Strassel, Mark Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. 2008. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *Proceedings of LREC*.

Jose L. Vicedo and Antonio Ferrndez. 2006. Coreference in Q&A. In Tomek Strzalkowski and Sanda M. Harabagiu, editors, *Advances in Open Domain Question Answering*, pages 71–96. Springer.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. 2009. A machine learning approach to textual entailment recognition. *Journal of Natural Language Engineering*, 15(4):551–582.

Dmitry Zelenko, Chinatsu Aone, and Jason Tibbetts. 2004. Coreference resolution for information extraction. In *Proceedings of the ACL Workshop on Reference Resolution and its Applications*.

Chapter 7

Recognising Entailment within Discourse

Recognising Entailment within Discourse

Shachar Mirkin[§], **Jonathan Berant**[†], **Ido Dagan**[§], **Eyal Shnarch**[§]

§ Computer Science Department, Bar-Ilan University
† The Blavatnik School of Computer Science, Tel-Aviv University

Abstract

Texts are commonly interpreted based on the entire discourse in which they are situated. Discourse processing has been shown useful for inference-based application; yet, most systems for textual entailment – a generic paradigm for applied inference – have only addressed discourse considerations via off-the-shelf coreference resolvers. In this paper we explore various discourse aspects in entailment inference, suggest initial solutions for them and investigate their impact on entailment performance. Our experiments suggest that discourse provides useful information, which significantly improves entailment inference, and should be better addressed by future entailment systems.

1 Introduction

This paper investigates the problem of recognising textual entailment within discourse. Textual Entailment (TE) is a generic framework for applied semantic inference (Dagan et al., 2009). Under TE, the relationship between a *text* (T) and a textual assertion (*hypothesis*, H) is defined such that *T entails H* if humans reading T would infer that H is most likely true (Dagan et al., 2006).

TE has been successfully applied to a variety of natural language processing applications, including information extraction (Romano et al., 2006) and question answering (Harabagiu and Hickl, 2006). Yet, most entailment systems have thus far paid little attention to discourse aspects of inference. In part, this is the result of the unavailability of adept tools for handling the kind of discourse processing required for inference. In addition in the main TE benchmarks, the Recognising Textual Entailment (RTE) challenges, discourse played little role. This state of affairs has started to change with the recent introduction of the RTE Pilot "Search" task (Bentivogli et al., 2009b), in which assessed texts are situated within complete documents. In this setting, texts need to be interpreted based on their entire discourse (Bentivogli et al., 2009a), hence attending to discourse issues becomes essential. Consider the following example from the task's dataset:

(T) *The seven men on board were said to have as little as 24 hours of air.*

For the interpretation of T, e.g. the identity and whereabouts of the seven men, one must consider T's discourse. The preceding sentence T', for instance, provides useful information to that aim:

(T') *The Russian navy worked desperately to save a small military submarine*.

This example demonstrates a common situation in texts, and is also applicable to the RTE Search task's setting. Still, little was done by the task's participants to consider discourse, and sentences were mostly processed independently.

Analyzing the Search task's development set, we identified several key discourse aspects that affect entailment in a discourse-dependent setting. First, we observed that the coverage of available coreference resolution tools is considerably limited. To partly address this problem, we extend the set of coreference relations to phrase pairs with a certain degree of lexical overlap, as long as no semantic incompatibility is found between them. Second, many bridging relations (Clark, 1975) are realized in the form of "global information" perceived as known for entire documents. As bridging falls completely out of the scope of available resolvers, we address this phenomenon by identifying and weighting prominent document terms and allowing their incorporation in inference even

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 770–778,
Beijing, August 2010

when they are not explicitly mentioned in a sentence. Finally, we observed a coherence-related discourse phenomenon, namely inter-relations between entailing sentences in the discourse, such as the tendency of entailing sentences to be adjacent to one another. To that end, we apply a two-phase classification scheme, where a second-phase meta-classifier is applied, extracting discourse and document-level features based on the classification of each sentence on its own.

Our results show that, even when simple solutions are employed, the reliance on discourse-based information is helpful and achieves a significant improvement of results. We analyze the contribution of each component and suggest some future work to better attend to discourse in entailment systems. To our knowledge, this is the most extensive effort thus far to empirically explore the effect of discourse on entailment systems.

2 Background

Discourse plays a key role in text understanding applications such as question answering or information extraction. Yet, such applications typically only handle a narrow aspect of discourse, addressing coreference by term substitution (Dali et al., 2009; Li et al., 2009). The limited coverage and scope of existing tools for coreference resolution and the unavailability of tools for addressing other discourse aspects also contribute to this situation. For instance, VP anaphora and bridging relations are usually not handled at all by such resolvers. A similar situation is seen in the TE research field.

The prominent benchmark for entailment systems evaluation is the series of RTE challenges. The main task in these challenges has traditionally been to determine, given a text-hypothesis pair (T,H), whether T entails H. Discourse played no role in the first two RTE challenges as T's were constructed of short simplified texts. In RTE-3 (Giampiccolo et al., 2007), where some paragraph-long texts were included, inter-sentential relations became relevant for correct inference. Yet the texts in the task were manually modified to ensure they are self-contained. Consequently, little effort was invested by the challenges' participants to address discourse issues beyond the standard substitution of coreferring nominal phrases, using publicly available tools such as JavaRap (Qiu et al., 2004) or OpenNLP[1], e.g. (Bar-Haim et al., 2008).

A major step in the RTE challenges towards a more practical setting of text processing applications occurred with the introduction of the Search task in the Fifth RTE challenge (RTE-5). In this task entailing sentences are situated within documents and depend on other sentences for their correct interpretation. Thus, discourse becomes a substantial factor impacting inference. Surprisingly, discourse hardly received any treatment in this task beyond the standard use of coreference resolution (Castillo, 2009; Litkowski, 2009), and an attempt to address globally-known information by removing from H words that appear in document headlines (Clark and Harrison, 2009).

3 The RTE Search Task

The RTE-5 Search task was derived from the TAC Summarization task[2]. The dataset consists of several corpora, each comprised of news articles concerning a specific *topic*, such as the impact of global warming on the Arctic or the London terrorist attacks in 2005. *Hypotheses* were manually generated based on Summary Content Units (Nenkova et al., 2007), clause-long statements taken from manual summaries of the corpora. *Texts* are unmodified sentences in the articles. Given a topic and a hypothesis, entailment systems are required to identify all sentences in the topic's corpus that entail the hypothesis.

Each sentence-hypothesis pair in both the development and test sets was annotated, judging whether the sentence entails the hypothesis. Out of 20,104 annotations in the development set, only 810 were judged as positive. This small ratio (4%) of positive examples, in comparison to 50% in traditional RTE tasks, better corresponds to the natural distribution of entailing texts in a corpus, thus better simulates practical settings.

The task may seem as a variant of information retrieval (IR), as it requires finding specific texts in a corpus. Yet, it is fundamentally different from IR for two reasons. First, the target output is a set

[1] http://opennlp.sourceforge.net

[2] http://www.nist.gov/tac/2009/Summarization/

of sentences, each evaluated independently, rather than a set of documents. Second, the decision criterion is *entailment* rather than *relevance*.

Despite the above, apparently, IR techniques provided hard-to-beat baselines for the RTE Search task (MacKinlay and Baldwin, 2009), outperforming every other system that relied on inference without IR-based pre-filtering. At the current state of performance of entailment systems, it seems that lexical coverage largely overshadows any other approach in this task. Still, most (6 out of 8) participants in the challenge applied their entailment systems to the entire dataset without a prior retrieval of candidate sentences. F_1 scores for such systems vary between 10% and 33%, in comparison to over 40% of the IR-based methods.

4 The Baseline RTE System

In this work we used BIUTEE, Bar-Ilan University Textual Entailment Engine (Bar-Haim et al., 2008; Bar-Haim et al., 2009), a state of the art RTE system, as a baseline and as a basis for our discourse-based enhancements. This section describes this system's architecture; the methods by which it was augmented to address discourse are presented in Section 5.

To determine entailment, BIUTEE performs the following main steps:

Preprocessing First, all documents are parsed and processed with standard tools for named entity recognition (Finkel et al., 2005) and coreference resolution. For the latter purpose, we use OpenNLP and enable the substitution of coreferring terms. This is the only way by which BIUTEE addresses discourse, representing the state of the art in entailment systems.

Entailment-based transformations Given a T-H pair (both represented as dependency parse trees), the system applies a sequence of knowledge-based entailment transformations over T, generating a set of texts which are entailed by it. The goal is to obtain consequent texts which are more similar to H. Based on preliminary results on the development set, in our experiments (Section 6) we use WordNet (Fellbaum, 1998) as the system's only knowledge resource, using its synonymy, hyponymy and derivation relations.

Classification A supervised classifier, trained on the development set, is applied to determine entailment of each pair based on a set of syntactic and lexical syntactic features assessing the degree by which T and its consequents cover H.

5 Addressing Discourse

In the following subsections we describe the prominent discourse phenomena that affect inference, which we have identified in an analysis of the development set and addressed in our implementation. As mentioned, these phenomena are poorly addressed by available reference resolvers or fall completely out of their scope.

5.1 Augmented coreference set

A large number of coreference relations are comprised of terms which share lexical elements, (e.g. *"airliners's first flight"* and *"Airbus A380's first flight"*). Although common in coreference relations, standard resolvers miss many of these cases. For the purpose of identifying additional coreferring terms, we consider two noun phrases in the same document as coreferring if: (i) their heads are identical and (ii) no semantic incompatibility is found between their modifiers. The types of incompatibility we handle are: (a) mismatching numbers, (b) antonymy and (c) co-hyponymy (coordinate terms), as specified by WordNet. For example, two nodes of the noun *distance* would be considered incompatible if one is modified by *short* and the second by its antonym *long*. Similarly, two modifier co-hyponyms of *distance*, such as *walking* and *running* would also result such an incompatibility. Adding more incompatibility types (e.g. *first* vs. *second flight*) may further improve the precision of this method.

5.2 Global information

Key terms or prominent pieces of information that appear in the document, typically at the title or the first few sentences, are many times perceived as "globally" known throughout the document. For example, the geographic location of the document theme, mentioned at the beginning of the document, is assumed to be known from that point on, and will often not be mentioned explicitly in further sentences. This is a bridging phenomenon

that is typically not addressed by available discourse processing tools. To compensate for that, we identify key terms for each document based on *tf-idf* scores and consider them as global information for that document. For example, global terms for the topic discussing the ice melting in the Arctic, typically contain a location such as *Arctic* or *Antarctica* and terms referring to *ice*, like *permafrost* or *iceshelf*.

We use a variant of *tf-idf*, where term frequency is computed as follows: $tf(t_{i,j}) = n_{i,j} + \vec{\lambda}^{\top} \cdot \vec{f_{i,j}}$. Here, $n_{i,j}$ is the frequency of term i in document j ($t_{i,j}$), which is incremented by additional positive weights ($\vec{\lambda}$) for a set of features ($\vec{f_{i,j}}$) of the term. Based on our analysis, we defined the following features, which correlated mostly with global information: (i) does the term appear in the title? (ii) is it a proper name? (iii) is it a location? The weights for these features are set empirically.

The document's top-n global terms are added to each of its sentences. As a result, a global term that occurs in the hypothesis is matched in each sentence of the document, regardless of whether the term explicitly appears in the sentence.

Considering the previous sentence Another method for addressing missing coreference and bridging relations is based on the assumption that adjacent sentences often refer to the same entities and events. Thus, when extracting classification features for a given sentence, in addition to the features extracted from the parse tree of the sentence itself, we extract the same set of features from the current and previous sentences together. Recall the example presented in Section 1. T is annotated as entailing the hypothesis *"The AS-28 mini-submarine was trapped underwater"*, but the word *submarine*, e.g., appears only in its preceding sentence T'. Thus, considering both sentences together when classifying T increases its coverage of the hypothesis. Indeed, a bridging reference relates *on board* in T with *submarine* in T', justifying our assumption in this case.

5.3 Document-level classification

Beyond discourse references addressed above, further information concerning discourse and document structure is available in the Search setting and may contribute to entailment classification. We observed that entailing sentences tend to come in bulks. This reflects a common coherence aspect, where the discussion of a specific topic is typically continuous rather than scattered across the entire document. This *locality* phenomenon may be useful for entailment classification since knowing that a sentence entails the hypothesis increases the probability that adjacent sentences entail the hypothesis as well.

To capture this phenomenon, we use a two-phase meta-classification scheme, in which a *meta-classifier* utilizes entailment classifications of the first classification phase to extract *meta-features* and determine the final classification decision. This scheme also provides a convenient way to combine scores from multiple classifiers used in the first classification phase. We refer to these as *base-classifiers*. This scheme and the meta-features we used are detailed hereunder.

Let us write (s, h) for a sentence-hypothesis pair. We denote the set of pairs in the development (training) set as $\mathcal{D}$ and in the test set as $\mathcal{T}$. We split $\mathcal{D}$ into two halves, $\mathcal{D}_1$ and $\mathcal{D}_2$. We make use of n base-classifiers, $C_1, \ldots, C_n$, among which $C^{\star}$ is a designated classifier with additional roles in the process, as described below. Classifiers may differ, for example, in their classification algorithm. An additional meta-classifier is denoted C_M. The classification scheme is shown as Algorithm 1.

Algorithm 1 Meta-classification

Training

1: Extract features for every (s, h) in $\mathcal{D}$
2: Train $C_1, \ldots, C_n$ on $\mathcal{D}_1$
3: Classify $\mathcal{D}_2$, using $C_1, \ldots, C_n$
4: Extract meta-features for $\mathcal{D}_2$ using the classification of $C_1, \ldots, C_n$
5: Train C_M on $\mathcal{D}_2$

Classification

6: Extract features for every (s, h) in $\mathcal{T}$
7: Classify $\mathcal{T}$ using $C_1, \ldots, C_n$
8: Extract meta-features for $\mathcal{T}$
9: Classify $\mathcal{T}$ using C_M

At Step 1, features are extracted for every (s, h) pair in the training set, as in the baseline system.

In Steps 2 and 3 we split the training set into two halves (taking half of each topic), train n different classifiers on the first half and then classify the second half using each of the n classifiers. Given the classification scores of the n base-classifiers to the (s, h) pairs in the second half of the training set, $\mathcal{D}_2$, we add in Step 4 the meta-features described in Section 5.3.1.

After adding the meta-features, we train (Step 5) a meta-classifier on this new set of features. Test sentences then go through the same process: features are extracted for them and they are classified by the already trained n classifiers (Steps 6 and 7), meta-features are extracted in Step 8, and a final classification decision is made by the meta-classifier in Step 9.

A retrieval step may precede the actual entailment classification, allowing the processing of fewer and potentially "better" candidates.

5.3.1 Meta-features

The following features are extracted in our meta-classification scheme:

Classification scores The classification score of each of the n base-classifiers.

Title entailment In many texts, and in news articles in particular, the title and the first few sentences often represent the entire document's content. Thus, knowing whether these sentences entail the hypothesis may be an indicator to the general potential of the document to include entailing sentences. Two binary features are added according to the classification of $C^\star$ indicating whether the title entails the hypothesis and whether the first sentence entails it.

Second-closest entailment Considering the locality phenomenon described above, we add a feature assigning higher scores to sentences in the vicinity of an entailment environment. This feature is computed as the distance to the second-closest entailing sentence in the document (counting the sentence itself as well), according to the classification of $C^\star$. Formally, let i be the index of the current sentence and $\mathcal{J}$ be the set of indices of entailing sentences in the document according to $C^\star$. For each $j \in \mathcal{J}$ we compute $d_{i,j} = |i-j|$, and choose the second smallest $d_{i,j}$ as d_i. The idea is that if entailing sentences indeed always come in bulks, then $d_i = 1$ for all entailing sentences, but $d_i > 1$ for all non-entailing ones. Figure 1 illustrates such a case, comparing the *second-closest* distance with the distance to the *closest* entailing sentence. In the *closest* scheme we do not count the sentence as closest to itself since it would disregard the environment of the sentence altogether, eliminating the desired effect. We scale the distance and add the feature score: $-\log(d_i)$.

#	Ent?	Closest	*d*	2nd closest	*d*
1	NO	6	5	7	6
2	NO	6	4	7	5
3	NO	6	3	7	4
4	NO	6	2	7	3
5	NO	6	1	7	2
6	YES	7	1	7	1
7	YES	6 or 8	1	6 or 8	1
8	YES	7 or 9	1	7 or 9	1
9	YES	8	1	8	1
10	NO	8	1	8	2
11	NO	8	2	8	3

Figure 1: Comparison of the *closest* and *second-closest* schemes when applied to a bulk of entailing sentences (in white) situated within a non-entailing environment (in gray). Unlike the *closest* one, the *second-closest* scheme assigns larger distance values to non-entailing sentences located on the 'edge' of the bulk (5 and 10) than to entailing ones.

Smoothed entailment This feature addressed the locality phenomenon by smoothing the classification score of sentence i with the scores of adjacent sentences, weighted by their distance from the current sentence i. Let $s(i)$ be the score assigned by $C^\star$ to sentence i. We add the Smoothed Entailment feature score:

$$\text{SE}(i) = \frac{\sum_w (b^{|w|} \cdot s(i+w))}{\sum_w (b^{|w|})}$$

where $0 < b < 1$ is the decay parameter and w is an integer bounded between $-N$ and N, denoting the distance from sentence *i*.

1st sentence entailing title Bensley and Hickl (2008) showed that the first sentence in a news article typically entails the article's title. We therefore assume that in each document, $s_1 \Rightarrow s_0$, where s_1 and s_0 are the document's first sentence and title respectively. Hence, under entailment transitivity, if $s_0 \Rightarrow h$ then $s_1 \Rightarrow h$. The corresponding binary feature states whether the sentence being classified is the document's first sentence *and* the title entails h according to $C^\star$.

	P (%)	R (%)	F_1 (%)
BIU-BL	14.53	55.25	23.00
BIU-DISC	20.82	57.25	30.53
BIU-BL^3	14.86	59.00	23.74
BIU-DISC$_{no-loc}$	22.35	57.12	32.13
All-yes baseline	4.6	100.0	8.9

Table 1: Micro-average results.

	P (%)	R (%)	F_1 (%)
By Topic			
BIU-BL	16.54	55.62	25.50
BIU-DISC	22.69	57.96	32.62
All-yes baseline	4.85	100.00	9.25
By Hypothesis			
BIU-BL	22.87	59.62	33.06
BIU-DISC	27.81	61.97	38.39
All-yes baseline	4.96	100.00	9.46

Table 2: Macro-average results.

Note that the above locality-based features rely on high accuracy of the base classifier $C^\star$. Otherwise, it will provide misleading information to the features computation. We analyze the effect of this accuracy in Section 6.

6 Results and Analysis

Using the RTE-5 Search data, we compare BIUTEE in its baseline configuration (cf. Section 4), denoted *BIU-BL*, with its discourse-aware enhancement (*BIU-DISC*) which uses all the components described in Section 5. To alleviate the strong IR effect described in Section 3, both systems are applied to the complete datasets (both training and test), without candidates pre-filtering.

BIU-DISC uses three base-classifiers ($n = 3$): *SVM*^{perf} (Joachims, 2006), and Naïve Bayes and Logistic Regression from the WEKA package (Witten and Frank, 2005). The first among these is set as our designated classifier $C^\star$, which is used for the computation of the document-level features. *SVM*^{perf} is also used for the meta-classifier. For the smoothed entailment score (cf. Section 5.3), we used $b = 0.9$ and $N = 3$. Global information is added by enriching each sentence with the highest-ranking term in the document, according to *tf-idf* scores (cf. Section 5.2), where document frequencies were computed based on about half a million documents from the TIPSTER corpus (Harman, 1992). The set of weights $\vec{\lambda}$ equals $\{2, 1, 4\}$ for title terms, proper names and locations, respectively. All parameters were tuned based on a 10-fold cross-validation on the development set, optimizing the micro-averaged F_1.

The results are presented in Table 1. As can be seen in the table, *BIU-DISC* outperforms *BIU-BL* in every measure, showing the impact of addressing discourse in this setting. To rule out the option that the improvement is simply due to the fact that we use three classifiers for *BIU-DISC* and a single one for *BIU-BL*, we show (*BIU-BL*^{3}) the results when the baseline system is applied in the same meta-classification configuration as *BIU-DISC*, with the same three classifiers. Apparently, without the discourse information this configuration's contribution is limited.

As mentioned in Section 5.3, the benefit from the locality features rely directly on the performance of the base classifiers. Hence, considering the low precision scores obtained here, we applied *BIU-DISC* to the data in the meta-classification scheme, but with locality features removed. The results, shown as *BIU-DISC*$_{no-loc}$ in the Table, indicate that indeed performance increases without these features. The last line of the table shows the results obtained by a naïve baseline where all test-set pairs are considered entailing.

For completeness, Table 2 shows the macro-averaged results, when averaged over the topics or over the hypotheses. Although we tuned our system to maximize micro-averaged F_1, these figures comply with the ones shown in Table 1.

Analysis of locality As discussed in Section 5, determining whether a sentence entails a hypothesis should take into account whether adjacent sentences also entail the hypothesis. In the above experiment we were unable to show the contribution of our system's component that attempts to capture this information; on the contrary, the results show it had a negative impact on performance.

Still, we claim that this information can be useful when used within a more accurate system. We try to validate this conjecture by understanding how performance of the locality features varies as the systems becomes more accurate. We do so via the following simulation.

When classifying a certain sentence, the classi-

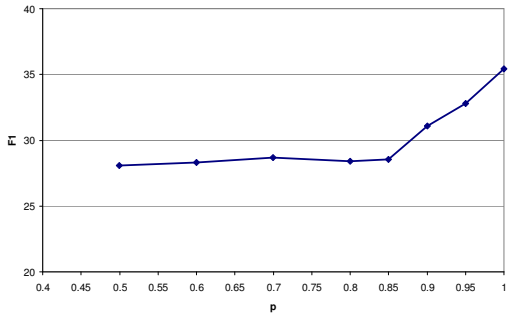

Figure 2: F_1 performance of *BIU-DISC* as a function of the accuracy in classifying adjacent sentences.

fications of its adjacent sentences are given by an *oracle classifier* that provides the correct answer with probability p. The system is applied using two locality features: the *1st sentence entailing title* feature and a close variant of the *smoothed entailment* feature, which calculates the weighted average of adjacent sentences, but disregards the score of the currently evaluated sentence.[3] Thus we supply information about adjacent sentences and test whether overall performance increases with the accuracy of this information.

We performed this experiment for p in a range of [0.5-1.0]. Figure 2 shows the results of this simulation, based on the average F_1 of five runs for each p. Since performance, from a certain point, increases with the accuracy of the oracle classifier, we can conclude that indeed precise information about adjacent sentences improves performance on the current sentence, and that locality is a true phenomenon in the data. We note, however, that performance improves only when accuracy is very high, suggesting the currently limited practical potential of this information, at least in the way locality was represented in this work.

Ablation tests Table 3 presents the results of the ablation tests performed to evaluate the contribution of each component. Based on the result reported in Table 1 and the above discussion, the tests were performed relative to *BIU-DISC*$_{no-loc}$, the optimal configuration. As seen in the table, the removal of each component causes a drop in results. For global information we see a minor difference, which is not surprising considering the conservative approach we took, using a single global term for each sentence. Possibly, this information is also included in the other components, thus proving no marginal contribution relative to them. Under the conditions of an overwhelming majority of negative examples, this is a risky method to use, and should be considered when the ratio of positive examples is higher. For future work, we intend to use this information via classification features (e.g. the coverage obtained with and without global information), rather than the crude addition of the term to the sentence.

[3]The *second-closest entailment* feature was not used as it considers the oracle's decision for the current sentence, while we wish to use only information about adjacent sentences.

Component removed	F_1 (%)	ΔF_1 (%)
Previous sent. features	28.55	3.58
Augmented coref.	26.73	5.40
Global information	31.76	0.37

Table 3: Results of ablation tests relative to *BIU-DISC*$_{no-loc}$. The columns specify the component removed, the micro-averaged F_1 score achieved without it, and the marginal contribution of the component.

Analysis of augmented coreferences We analyzed the performance of the component for augmenting coreference relations relative to the OpenNLP resolver. Recall that our component works on top of the resolver's output and can add or remove coreference relations. As a complete annotation of coreference chains in the dataset is unavailable, we performed the following evaluation. Recall is computed based on the number of identified pairs from a sample of 100 intra-document coreference and bridging relations from the annotated dataset described in (Mirkin et al., 2010). Precision is computed based on 50 pairs sampled from the output of each method, equally distributed over topics. The results, shown in Table 4, indicate the much higher recall obtained by our component at some cost in precision. Although rather simple, the ablation test of this component shows its usefulness. Still, both methods achieve low absolute recall, suggesting the need for more robust tools for this task.

	P (%)	R (%)	F_1 (%)
OpenNLP	74	16	26.3
Augmented coref.	60	28	38.2

Table 4: Performance of coreference methods.

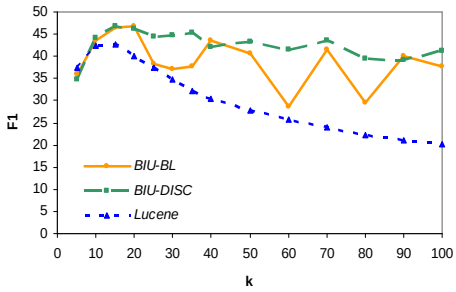

Figure 3: F_1 performance as a function of the number of retrieved candidates.

Candidate retrieval setting As mentioned in Section 3, best performance of RTE systems in the task was obtained when applying a first step of IR-based candidate filtering. We therefore compare the performance of *BIU-DISC* with that of *BIU-BL* under this setting as well.[4] For candidate retrieval we used Lucene, a state of the art search engine[5], in a range of top-k retrieved candidates. The results are shown in Figure 3. For reference, the figure also shows the performance along this range of Lucene as-is, when no further inference is applied to the retrieved candidates.

While *BIU-DISC* does not outperform *BIU-BL* at every point, the area under the curve is clearly larger for *BIU-DISC*. The figure also indicates that *BIU-DISC* is far more robust, maintaining a stable F_1 and enabling a stable tradeoff between recall and precision along the whole range (recall ranges between 42% and 55% for $k \in [15 - 100]$, with corresponding precision range of 51% to 33%).

Finally, Table 5 shows the results of the best systems as determined in our first experiment. We performed a single experiment to compare *BIU-DISC*$_{no-loc}$ and *BIU-BL*^{3} under a candidate retrieval setting, using $k = 20$, where both systems highly perform. We compare these results to the highest score obtained by Lucene, as well as to the two best submissions to the RTE-5 Search task[6]. *BIU-DISC*$_{no-loc}$ outperforms all other methods and its result is significantly better than *BIU-BL*^{3} with $p < 0.01$ according to McNemar's test.

[4]This time, for global information, the document's three highest ranking terms were added to each sentence.

[5]http://lucene.apache.org

[6]The best one is an earlier version of this work (Mirkin et al., 2009); the second is MacKinlay and Baldwin's (2009).

	P (%)	R (%)	F_1 (%)
BIU-DISC$_{no-loc}$	50.77	45.12	47.78
BIU-BL^3	51.68	40.38	45.33
Lucene, top-15	35.93	52.50	42.66
RTE-5 best	40.98	51.38	45.59
RTE-5 second-best	42.94	38.00	40.32

Table 5: Performance of best configurations.

7 Conclusions

While it is generally assumed that discourse interacts with semantic entailment inference, the concrete impacts of discourse on such inference have been hardly explored. This paper presented a first empirical investigation of discourse processing aspects related to entailment. We argue that available discourse processing tools should be substantially improved towards this end, both in terms of the phenomena they address today, namely nominal coreference, and with respect to the covering of additional phenomena, such as bridging anaphora. Our experiments show that even rather simple methods for addressing discourse can have a substantial positive impact on the performance of entailment inference. Concerning the locality phenomenon stemming from discourse coherence, we learned that it does carry potentially useful information, which might become beneficial in the future when better-performing entailment systems become available. Until then, integrating this information with entailment confidence may be useful. Overall, we suggest that entailment systems should extensively incorporate discourse information, while developing sound algorithms for addressing various discourse phenomena, including the ones described in this paper.

Acknowledgements

The authors are thankful to Asher Stern and Ilya Kogan for their help in implementing and evaluating the augmented coreference component, and to Roy Bar-Haim for useful advice concerning this paper. This work was partially supported by the Israel Science Foundation grant 1112/08 and the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886. Jonathan Berant is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship.

References

Bar-Haim, Roy, Jonathan Berant, Ido Dagan, Iddo Greental, Shachar Mirkin, Eyal Shnarch, and Idan Szpektor. 2008. Efficient semantic deduction and approximate matching over compact parse forests. In *Proc. of Text Analysis Conference (TAC)*.

Bar-Haim, Roy, Jonathan Berant, and Ido Dagan. 2009. A compact forest for scalable inference over entailment and paraphrase rules. In *Proc. of EMNLP*.

Bensley, Jeremy and Andrew Hickl. 2008. Unsupervised resource creation for textual inference applications. In *Proc. of LREC*.

Bentivogli, Luisa, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2009a. Considering discourse references in textual entailment annotation. In *Proc. of the 5th International Conference on Generative Approaches to the Lexicon (GL2009)*.

Bentivogli, Luisa, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009b. The fifth PASCAL recognizing textual entailment challenge. In *Proc. of TAC*.

Castillo, Julio J. 2009. Sagan in TAC2009: Using support vector machines in recognizing textual entailment and TE search pilot task. In *Proc. of TAC*.

Clark, Peter and Phil Harrison. 2009. An inference-based approach to recognizing entailment. In *Proc. of TAC*.

Clark, Herbert H. 1975. Bridging. In Schank, R. C. and B. L. Nash-Webber, editors, *Theoretical issues in natural language processing*, pages 169–174. Association of Computing Machinery.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Dagan, Ido, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, pages 15(4):1–17.

Dali, Lorand, Delia Rusu, Blaz Fortuna, Dunja Mladenic, and Marko Grobelnik. 2009. Question answering based on semantic graphs. In *Proc. of the Workshop on Semantic Search (SemSearch 2009)*.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of ACL*.

Giampiccolo, Danilo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Harabagiu, Sanda and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proc. of ACL*.

Harman, Donna. 1992. The DARPA TIPSTER project. *SIGIR Forum*, 26(2):26–28.

Joachims, Thorsten. 2006. Training linear SVMs in linear time. In *Proc. of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Li, Fangtao, Yang Tang, Minlie Huang, and Xiaoyan Zhu. 2009. Answering opinion questions with random walks on graphs. In *Proc. of ACL-IJCNLP*.

Litkowski, Ken. 2009. Overlap analysis in textual entailment recognition. In *Proc. of TAC*.

MacKinlay, Andrew and Timothy Baldwin. 2009. A baseline approach to the RTE5 search pilot. In *Proc. of TAC*.

Mirkin, Shachar, Roy Bar-Haim, Jonathan Berant, Ido Dagan Eyal Shnarch, Asher Stern, and Idan Szpektor. 2009. Addressing discourse and document structure in the RTE search task. In *Proc. of TAC*.

Mirkin, Shachar, Ido Dagan, and Sebastian Padó. 2010. Assessing the role of discourse references in entailment inference. In *Proc. of ACL*.

Nenkova, Ani, Rebecca Passonneau, and Kathleen Mckeown. 2007. The pyramid method: incorporating human content selection variation in summarization evaluation. In *ACM Transactions on Speech and Language Processing*.

Qiu, Long, Min-Yen Kan, and Tat-Seng Chua. 2004. A public reference implementation of the RAP anaphora resolution algorithm. In *Proc. of LREC*.

Romano, Lorenza, Milen Kouylekov, Idan Szpektor, and Ido Dagan. 2006. Investigating a generic paraphrase-based approach for relation extraction. In *Proc. of EACL*.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco.

Chapter 8

Conclusions and Discussion

This thesis addressed several aspects of applied semantic inference within the Textual Entailment framework, focusing on the impact of contextual information on inference. With the high-level goal of improving Textual Entailment inference, we have inspected and addressed several kinds of components that take part in the inference process and types of information that may be relevant for it. Through this work, we have provided insights and directions for the development of components useful for Textual Entailment modeling on one hand, and the incorporation of such components into entailment systems, on the other hand. While we employed Textual Entailment for inference modeling throughout this work, we suggest that a great deal of the ideas presented here are applicable to other semantic inference approaches. The following paragraphs recap the main contributions of this work. Section 8.1 suggests directions for future research for each of the topics addressed in this thesis.

In Chapter 3 we suggested a novel methodology for evaluating the utility of lexical-semantic resources for inference, based on operational definitions for lexical entailment that we established. We applied our methodology to several resources, presenting a comparable quantitative evaluation. Our evaluation points out the limits and gaps in existing resources and provides insights with respect to the potential improvements

in the utilization of existing resources and the development of new resources which are better-suited for Textual Entailment modeling. Notably, application of rules in inappropriate contexts was found to be an outstanding cause for the degradation of the practical utility of lexical resources.

The contribution of context validation to lexical inferences was shown in Chapter 4, where we addressed the problem of unknown terms in Machine Translation. We suggested using monolingual source-language information for generating alternative texts to the source sentence for translation, where prior works relied on multilingual information for that purpose. Further extending prior work, we proposed generating entailed sentences when paraphrasing was not possible, thus expanding the coverage of the MT system while still obtaining useful translations. Using context matching models, we ranked the modifications to the source sentence with respect to their suitability for the given context. Such ranking provides more appropriate translations and allows the process to be conducted more efficiently. Our results were significantly favorable in comparison with prior work.

In Chapter 5 we proposed a scheme to comprehensively address context matching in inference scenarios, which goes beyond the validation of rule applications as carried out in Chapter 4. This scheme provides a unified, directional approach for the context matching task and a natural way to incorporate various types of contextual information. An implementation of the scheme and its application to a Text Categorization task achieved state of the art results, suggesting its usage for other inference-based tasks.

In Chapters 6 and 7 we considered a different contextual aspect of text – discourse. We first (Chapter 6) analyzed and assessed the potential contribution of discourse references to inference. Our study provides insights and operational directions to developers of reference resolvers and of inference systems. To the former, by highlighting which types of references that are currently neglected by state of the art

systems affect inference; to the latter, by suggesting ways in which such information can be incorporated within entailment systems. Next (Chapter 7), we undertook the task of practically showing that discourse information can significantly improve the performance of inference systems if utilized, leading to a discourse-enhanced system that outperformed the state of the art.

In sum, inferences conducted by semantic inference systems often disregard the context surrounding the specific text fragments that participate in the inference operation. In this work we have shown that context, in its different aspects, is a major issue to consider, and when regarded, inference performance is substantially improved. Much can be done to further utilize the full potential provided by contextual information. Some directions are presented in the next section.

8.1 Future research

Generally speaking, in this work we addressed topics that were relatively unexplored previously. We strived to provide a better understanding of the impact of various aspects, components or subtasks of Textual Entailment modeling, through evaluation, analysis or empirical studies. Our research revealed many directions that require deeper exploration. The main ones among them are detailed in this section. We conclude by elaborating on the potential usages of Textual Entailment to Machine Translation, as a continuation to the utilization of Textual Entailment for Machine Translation, which was presented in Chapter 4.

Lexical entailment

Our analysis of existing lexical-semantic resources (Chapter 3) reveals current gaps in knowledge resources for entailment-based inference, in particular the unavailability of enough dedicated resources for entailment rules. As a result, in order to obtain

more coverage, entailment rules are indirectly derived from resources which are often non-directional or provide fuzzy evidence in terms of entailment, namely *similarity*. Since the publication of the work presented in Chapter 3, several resources have been released, but the gap is still significant. As mentioned, the practical utility of resources is degraded by the fact that rules are used out of context. Developing context matching methods, such as the one suggested in Chapter 5, is required to address the problem on the inference system side (and see more below), but more can be done on the resources side as well. Available lexical resources are still missing information that may be helpful for inference purposes. For example, a resource that contains the corpus instances based on which rules were learned may make it easier for a rule context classifier to determine whether the given context, to which the rule is to be applied, is suitable for the rule. Rule scores may be valuable as well and may serve as a prior indicator to favor either more commonly applicable rules or rules learned with higher confidence.

Various types of lexical-semantic relations may be useful for entailment inference, from the most commonly used is-a relation to rules encoding complex world knowledge. We have shown that substitutable relations are not sufficient and that non-substitutable relations may significantly contribute to recall. It is easy to grasp the importance of these relationships for inference when considering Text Categorization tasks, as in Chapter 5. For instance, the rule '*excavation* $\Rightarrow$ *archeology*' may be helpful for categorizing the document in which *excavation* occurs to the *archeology* category, and may be regarded as a case of "topical" entailment. Nonetheless, it is still widely unknown what types of "non-standard" relations conform with Textual Entailment. A typology of such relations may help clarify the picture, and may contribute to the development of acquisition methods for entailment rules that extend beyond the relations that are regularly extracted. Employing non-substitutable rules would require an effort also on the inference systems' side in order to accommodate

the application of such rules.

Our work showed the benefit of using distributional similarity resources for entailment, as they increase rule coverage. Yet, such resources contain many incorrect rules from the entailment point of view, such as pairs of antonyms or co-hyponyms which are frequently contextually-similar. Automatically filtering rules from such resources may make the resources more useable for Textual Entailment modeling needs.

A different direction is to learn "dynamic" entailment rules from the test text. Available knowledge is limited, yet much of the knowledge required for inference can be found in the text itself. Such knowledge is frequently context-specific and cannot always be included in general-purpose knowledge resources; instead, it needs to be acquired on-the-fly. This can be addressed by developing methods for extracting such knowledge from the current text and using the currently-learned knowledge within the text's own scope. For instance, from Example 8.1a we can learn the rules '*deep-water soloing* $\Rightarrow$ *rock climbing*' and '*DWS* $\Leftrightarrow$ *deep-water soloing*', which are applicable in the context of rock climbing. These rules can assist in inferring the hypothesis "*Chris Sharma climbed in Spain*" from 8.1b.

(8.1) (a) *Considered by many to be the best rock climber in the world, Chris Omprakash Sharma is originally from Santa Cruz, CA. [. . .] Sharma now focuses on three types of rock climbing: bouldering, sport climbing, and deep-water soloing (DWS).*

(b) *In September 2006, with well over 80 attempts made, Sharma finally completed his 20m DWS arch route off the east coast of the island of Mallorca (Spain).*

Additional rules can be learned from the above texts during inference (e.g. '*Mallorca* $\Rightarrow$ *island*' and '*Mallorca* $\Rightarrow$ *Spain*') and may be added, under certain conditions, to the general-purpose rulebase for future use. As mentioned in Chapter 6, discourse

relationships may also be of use for this purpose, providing another source of knowledge to complement the "static" knowledge resources.

Context matching

In Chapter 5 we presented a generic scheme for Contextual Preferences and demonstrated it using a Text Categorization task. To the best of our knowledge, ours is the only implementation of CP apart from the methods suggested along with the original proposal of the CP framework in (Szpektor et al., 2008). We hope that our work will encourage further research about comprehensively addressing context matching, where all aspects of inference are regarded. Specifically, applying this scheme to other inference-based applications is desired, in particular to tasks where more complex hypotheses are involved. Another interesting research topic is the application of the classification-based approach to settings where hypotheses are not provided in advance, but are rather given online, such as in the case of Information Retrieval, where queries play the role of hypotheses.

At the same time, current entailment systems do not attend to context issues for the most part. Context matching models should be embedded within inference systems to ensure that matches are context-validated before relying on them for inference.

An open issue with methods like ours, which train classifiers for each hypothesis or each rule, is of scalability, especially when hypotheses are unknown ahead of time and the number of potential rules is large. One direction is to develop efficient methods for training classifiers on-the-fly. Another option is to train rule classifiers for entire resources in advance, especially since scalability concerns rule classifiers more than hypothesis classifiers, simply due to the larger volume of rule classifiers that need to be trained. Another interesting direction is training a global classifier, in the spirit of Connor and Roth (2007), or training a reduced set of classifiers, which are suitable

for sets of words rather than a single one.

Discourse

Inference systems, and entailment systems in particular, have largely overlooked the information available through the complete provided text and have handled only a narrow aspect of discourse. Our work suggests that the discourse provides many types of information that are useful for inference.

Among other things, we have shown that a considerable number of discourse references are comprised of verbal phrases and bridging relations. We have further shown that quite simple semantic augmentations of coreference tools may be helpful in increasing their coverage. In addition, other discourse-related phenomena such as coherence or topic continuity are apparently relevant and may contribute to inference performance if addressed. Accommodating to these needs requires an effort on the discourse processing side, in order to develop tools which will capture more types of discourse relations and phenomena than are addressed today. Adapting entailment systems to utilize such information is necessary as well. Specifically, only substitution operations are currently supported in entailment systems, while other operations, such as *merge* and *insertion* which were presented in Chapter 6, are needed in order to accommodate to the different types of relevant discourse relations. Other discourse phenomena motivate further adaptation of systems in order to exploit the information available in the discourse. Beyond the ones discussed in this work, identifying additional types of discourse information that can contribute to inference is another direction worth investigating.

An additional topic for future research concerns the interplay between discourse references and entailment knowledge (see Chapter 6). Our analysis suggests that reference resolution constitutes a source of context-specific knowledge that is complementary to generic entailment knowledge. While knowledge was previously used for

reference detection, reference relations were not yet utilized to reassure or complement entailment rules. The two information sources may be used to reinforce each other. Specifically, discourse references may be employed to validate the applicability of entailment rules.

A challenging task for future work, which concerns all of the above topics, is the design of interfaces between entailment systems and the various components they use, including knowledge resources, context matching models and discourse processing tools. We believe that a reasonably simple integration scheme would encourage the adaptation of tools and resources for inference tasks. Further, being able to easily use one's favorite tool or apply the most suitable one for the task will encourage the use of inference systems in general and is expected to improve inference performance as well.

Textual Entailment for Machine Translation

In Chapter 4 we demonstrated how Textual Entailment can be used to address the problem of out-of-vocabulary terms in Machine Translation through the generation and translation of sentences that are entailed by the source sentence. In (Aziz et al., 2010) we further pursue this idea, while improving on the way entailment rules are integrated into the MT flow. Other works (Callison-Burch et al., 2006; Marton et al., 2009) utilized paraphrases for this task, thus improving the overall performance of their MT systems. Additionally, Kauchak and Barzilay (2006) used paraphrasing to overcome language variability in Machine Translation Evaluation, and Padó et al. (2009) presented the relatedness of the Machine Translation Evaluation and Textual Entailment recognition tasks and used Textual Entailment features to gain improved MT Evaluation.

We suggest that monolingual Textual Entailment technology can be utilized more

extensively in various steps throughout the MT flow. First, when the system cannot translate the source sentence or when it does so with low confidence, entailment transformations may be employed to modify the original sentence by replacing unknown terms with equivalent or entailed ones, as well as by simplifying or removing text constructs by using syntactic entailment rules. Unlike arbitrary removal of text portions, using Textual Entailment for this task ensures (to the extent of the Textual Entailment system's accuracy) that the modifications are principled, maintaining an equivalence or entailment relation between the original and the modified texts. For example, an entailment operation may be the removal of a clause or a modifier that are making the translation more difficult. For a concrete example consider Example 8.2 below. Each of the boldface expressions in the original sentence 8.2a can be removed or modified, based on the MT system's needs, to generate an entailed sentence 8.2b that is easier for the system at hand to translate, consequently producing a potentially better translation (see the discussion later about the information that is lost through this procedure).

(8.2) (a) *Uma Thurman* ***starred*** *in Pulp Fiction,* ***one of several films*** *directed by* ***the renowned*** *Quentin Tarantino.*

(b) *Uma Thurman acted in Pulp Fiction, directed by Quentin Tarantino.*

Second, once a translation had been generated, Textual Entailment technology may be used to generate an equivalent or entailed sentence in the target language with improved fluency, beyond the local capability of the MT system's target language model. Lastly, the paraphrasing approach for Machine Translation Evaluation by Kauchak and Barzilay (2006) can be extended to directional entailment by considering entailment relations between reference translations and the MT system's output. If a reference directionally entails the system's output, then the translation may still be useful, albeit missing some information, as in Example 8.3 where English is the

target language. In comparison with Padó et al. (2009) who preserved the original MT Evaluation goal of assessing equivalence and thus compared the output and the reference in both directions, we specifically point at the entailed cases of reference-output pairs which may be the result of the steps for using TE for MT suggested above.

(8.3) Reference: *The Tunisian embassy in Switzerland was hit by firebombs.*

MT output: *The embassy of Tunisia in Switzerland was attacked.*

An issue that should be taken into account when applying directional entailment transformations for Machine Translation is that some of the information that was originally conveyed in the source sentence may be lost. It might be useful to know which information can be discarded and to measure how much information is lost by such operations. This can enable deciding which operations can be applied to the text, determining the confidence of the translation and providing an indication to the user as to the magnitude of difference between the original and translated texts. This is a standing challenge.

In summary, we suggest that there are several benefits of integrating Textual Entailment technology within the MT flow. Our work constitutes merely the first steps in such directions, leaving much room for future research.

Bibliography

Adams, R., Nicolae, G., Nicolae, C., and Harabagiu, S. (2007). Textual Entailment through Extended Lexical Overlap and Lexico-semantic Matching. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 119–124, Prague, Czech Republic.

Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability Assessment for Text Simplification. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*, Los Angeles, CA, USA.

Androutsopoulos, I. and Malakasiotis, P. (2009). A Survey of Paraphrasing and Textual Entailment Methods. *CoRR*, abs/0912.3747.

Asher, N. and Lascarides, A. (1998). Bridging. *Journal of Semantics*, 15(1):83–113.

Aziz, W., Dymetmany, M., Mirkin, S., Specia, L., Cancedda, N., and Dagan, I. (2010). Learning an Expert from Human Annotations in Statistical Machine Translation: the Case of Out-of-VocabularyWords. In *Proceedings of the 14th annual meeting of the European Association for Machine Translation (EAMT)*, Saint-Raphaël, France.

Balahur, A., Lloret, E., Óscar Ferrández, Montoyo, A., Palomar, M., and Muñoz, R. (2008). The DLSIUAES Team's Participation in the TAC 2008 Tracks. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Bar-Haim, R., Berant, J., and Dagan, I. (2009). A Compact Forest for Scalable Inference over Entailment and Paraphrase Rules. In *Proceedings of EMNLP*, Suntec, Singapore.

Bar-Haim, R., Berant, J., Dagan, I., Greental, I., Mirkin, S., and amd Idan Szpektor, E. S. (2008a). Efficient Semantic Deduction and Approximate Matching over Compact Parse Forests. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Bar-Haim, R., Berant, J., Dagan, I., Greental, I., Mirkin, S., Eyal, S., and Szpektor, I. (2008b). Efficient Semantic Deduction and Approximate Matching over Compact Parse Forests. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Bar-Haim, R., Berant, J., Dagan, I., Greental, I., Mirkin, S., Shnarch, E., and Szpektor, I. (2008c). Efficient Semantic Deduction and Approximate Matching over Compact Parse Forests. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., and Szpektor, I. (2006). The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenge Workshop for Recognising Textual Entailment*, Venice, Italy.

Bar-Haim, R., Dagan, I., Greental, I., and Shnarch, E. (2007). Semantic Inference at the Lexical-Syntactic Level. In *Proceedings of AAAI*, pages 871–870, Vancouver, British Columbia, Canada.

Barak, L., Dagan, I., and Shnarch, E. (2009). Text Categorization from Category Name via Lexical Reference. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 33–36, Boulder, Colorado, USA.

Bejan, C. A. and Harabagiu, S. (2010). Unsupervised Event Coreference Resolution with Rich Linguistic Features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden.

Ben-Aharon, R., Szpektor, I., and Dagan, I. (2010). Generating Entailment Rules from FrameNet. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 241–246, Uppsala, Sweden.

Bentivogli, L., Clarck, P., Dagan, I., Dang, H. T., and Giampiccolo, D. (2010). The Sixth PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Bentivogli, L., Dagan, I., Dang, H. T., Giampiccolo, D., and Magnini, B. (2009). The Fifth PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Berant, J., Dagan, I., and Goldberger, J. (2011). Global Learning of Typed Entailment Rules. In *Proceedings of the 49th Annual Meeting of the Association for*

Computational Linguistics: Human Language Technologies, pages 610–619, Portland, Oregon, USA.

Berland, M. and Charniak, E. (1999). Finding Parts in Very Large Corpora. In *Proceedings of ACL*, pages 57–64, College Park, Maryland, USA.

Bos, J. and Markert, K. (2005). Recognising Textual Entailment with Logical Inference. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 628–635, Vancouver, British Columbia, Canada.

Brin, S. (1998). Extracting Patterns and Relations from the World Wide Web. In *Proceedings of WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT98*, pages 172–183, Valencia, Spain.

Burchardt, A., Pennacchiotti, M., Thater, S., and Pinkal, M. (2009). Assessing the Impact of Frame Semantics on Textual Entailment. *Journal of Natural Language Engineering*, 15(4):527–550.

Burstein, J., Marcu, D., and Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, 18:32–39.

Cabrio, E., Kouylekov, M., and Magnini, B. (2008). Combining Specialized Entailment Engines for RTE-4. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithsburg, Maryland, USA.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of HLT-NAACL*, New York City, New York, USA.

Carpuat, M. and Wu, D. (2007). Improving Statistical Machine Translation Using Word Sense Disambiguation. In *In The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007*, pages 61–72, Prague, Czech Republic.

Castilloa, J. J. and i Alemany, L. A. (2008). An approach using Named Entities for Recognizing Textual Entailment. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithsburg, Maryland, USA.

Chai, J. Y. and Jin, R. (2004). Discourse Structure for Context Question Answering. In *HLT-NAACL 2004: Workshop on Pragmatics of Question Answering*, pages 23–30, Boston, Massachusetts, USA.

Chierchia, G. and McConnell-Ginet, S. (2000). *Meaning and Grammar: an Introduction to Semantics*. MIT Press.

Chklovski, T. and Pantel, P. (2004). VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain.

Clark, H. H. (1975). Bridging. In Schank, R. C. and Nash-Webber, B. L., editors, *Theoretical Issues in Natural Language Processing*, pages 169–174. Association of Computing Machinery.

Clark, P. and Harrison, P. (2008). Recognizing Textual Entailment with Logical Inference. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithsburg, Maryland, USA.

Clark, P. and Harrison, P. (2010). BLUE-Lite: A Knowledge-Based Lexical Entailment System for RTE6. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Clarke, J. and Lapata, M. (2010). Discourse Constraints for Document Compression. *Computational Linguistics*, 36(3):411–441.

Connor, M. and Roth, D. (2007). Context Sensitive Paraphrasing with a Global Unsupervised Classifier. In *Proceedings of the 18th European Conference on Machine Learning (ECML)*, pages 104–115, Warsaw, Poland.

Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2009). Recognizing Textual Entailment: Rational, Evaluation and Approaches. *Natural Language Engineering*, pages 15(4):1–17.

Dagan, I. and Glickman, O. (2004). Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In *PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France.

Dagan, I. and Glickman, O. (2005). The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, Southampton, UK.

Dagan, I., Glickman, O., Gliozzo, A., Marmorshtein, E., and Strapparava, C. (2006a). Direct Word Sense Matching for Lexical Substitution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 449–456, Sydney, Australia.

Dagan, I., Glickman, O., and Magnini, B. (2006b). The PASCAL Recognising Textual Entailment Challenge. *Lecture Notes in Computer Science*, 3944:177–190.

de Marneffe, M.-C., Rafferty, A. N., and Manning, C. D. (2008). Finding Contradictions in Text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio, USA.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of HLT*, San Diego, California, USA.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Galanis, D. and Malakasiotis, P. (2008). AUEB at TAC 2008. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Giampiccolo, D., Dang, H. T., Magnini, B., Dagan, I., Cabrio, E., and Dolan, B. (2008). The Forth PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007). The Third PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic.

Glickman, O., Shnarch, E., and Dagan, I. (2006). Lexical Reference: a Semantic Matching Subtask. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 172–179, Sydney, Australia.

Habash, N. and Dorr, B. (2003). A Categorial Variation Database For English. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL '03, pages 17–23, Edmonton, Canada.

Harabagiu, S. and Hickl, A. (2006). Methods for Using Textual Entailment in Open-domain Question Answering. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 905–912, Sydney, Australia.

Harabagiu, S., Hickl, A., and Lacatusu, F. (2007). Satisfying Information Needs with Multi-document Summaries. *Inf. Process. Manage.*, 43:1619–1642.

Harabagiu, S. M., Maiorano, S. J., and Paşca, M. A. (2003). Open-domain Textual Question Answering Techniques. *Natural Language Engineering*, 9(3):231–267.

Harmeling, S. (2009). Inferring Textual Entailment with a Probabilistically Sound Calculus. *Journal of Natural Language Engineering*, pages 459–477.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (Coling)*, pages 539–545, Nantes, France.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% Solution. In *Proceedings of HLT-NAACL 2006*, pages 57–60, New York City, New York, USA.

Hughes, T. and Ramage, D. (2007). Lexical Semantic Relatedness with Random Graph Walks. In *Proceedings of EMNLP-CoNLL*, pages 581–589, Prague, Czech Republic.

Humphreys, K., Gaizauskas, R., and Azzam, S. (1997). Event Coreference for Information Extraction. In *Proceedings of the Workshop on Operational Factors In Practical, Robust Anaphora Resolution for Unrestricted Texts, 35Th Meeting Of ACL*, pages 75–81, Madrid, Spain.

Iftene, A. and Balahur-Dobrescu, A. (2007). Hypothesis Transformation and Semantic Variability Rules Used in Recognizing Textual Entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 125–130, Prague, Czech Republic.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). Text Simplification for Reading Assistance: a Project Note. In *Proceedings of the 2nd International Workshop on Paraphrasing*, PARAPHRASE '03, pages 9–16, Sapporo, Japan.

Kauchak, D. and Barzilay, R. (2006). Paraphrasing for Automatic Evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, New York, USA.

Kipper, K., Dang, H. T., and Palmer, M. (2000). Class-Based Construction of a Verb Lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI)*, pages 691–696, Austin, Texas, USA.

Kotlerman, L., Dagan, I., Szpektor, I., and Zhitomirsky-Geffet, M. (2010). Directional Distributional Similarity for Lexical Inference. *Natural Language Engineering*, 16(4):359–389.

Kouylekov, M. and Magnini, B. (2005). Recognizing Textual Entailment with Tree Edit Distance Algorithms. In *Proceedings of the PASCAL Recognising Textual Entailment Challenge*, pages 17–20, Southampton, UK.

Kouylekov, M. and Magnini, B. (2006). Tree Edit Distance for Recognizing Textual Entailment: Estimating the Cost of Insertion. In *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge*, pages 68–73, Venice, Italy.

Lavie, A. and Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic.

Lin, D. (1998a). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 768–774, Montreal, Quebec, Canada.

Lin, D. (1998b). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING-ACL*, Montreal, Quebec, Canada.

Lin, D. (1998c). Dependency-based evaluation of MINIPAR. In *Proceedings of the Workshop on Evaluation of Parsing Systems at LREC 1998*, Granada, Spain.

Lin, D. and Pantel, P. (2001). Discovery of Inference Rules for Question Answering. *Natural Language Engineering*, 7(4):343–360.

Lindn, K. and Piitulainen, J. (2004). Discovering Synonyms and Other Related Words. In *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*, pages 63–70, Geneva, Switzerland.

MacCartney, B., Galley, M., and Manning, C. D. (2008). A Phrase-based Alignment Model for Natural Language Inference. In *Proceedings of EMNLP*, pages 802–811, Honolulu, Hawaii.

Macleod, C., Grishman, R., Meyers, A., Barrett, L., and Reeves, R. (1998). NOMLEX: A Lexicon of Nominalizations. In *Proceedings of Euralex98*, pages 187–193, Liege, Belgium.

Markert, K. and Nissim, M. (2003). Using the Web for Nominal Anaphora Resolution. In *In EACL Workshop on the Computational Treatment of Anaphora*, pages 39–46, Budapest, Hungary.

Marsi, E., Krahmer, E., Bosma, W., and Theune, M. (2006). Normalized Alignment of Dependency Trees for Detecting Textual Entailment. In *The Second PASCAL Recognising Textual Entailment Challenge*, pages 56–61, Venice, Italy.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. In *Proceedings of EMNLP*, pages 381–390, Suntec, Singapore.

McCarthy, D. and Navigli, R. (2009). The English Lexical Substitution Task. *Language Resources and Evaluation*, 43(2):139–159.

Mirkin, S., Bar-Haim, R., Berant, J., Dagan, I., Shnarch, E., Stern, A., and Szpektor, I. (2009a). Addressing Discourse and Document Structure in the RTE Search Task. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Mirkin, S., Berant, J., Dagan, I., and Shnarch, E. (2010a). Recognising Entailment within Discourse. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 770–778, Beijing, China.

Mirkin, S., Dagan, I., and Geffet, M. (2006). Integrating Pattern-Based and Distributional Similarity Methods for Lexical Entailment Acquisition. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 579–586, Sydney, Australia.

Mirkin, S., Dagan, I., Kotlerman, L., and Szpektor, I. (2011). Classification-based Contextual Preferences. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 20–29, Edinburgh, Scotland, UK.

Mirkin, S., Dagan, I., and Padó, S. (2010b). Assessing the Role of Discourse References in Entailment Inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1209–1219, Uppsala, Sweden.

Mirkin, S., Dagan, I., and Shnarch, E. (2009b). Evaluating the Inferential Utility of Lexical-Semantic Resources. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 558–566, Athens, Greece.

Mirkin, S., Specia, L., Cancedda, N., Dagan, I., Dymetman, M., and Szpektor, I. (2009c). Source-Language Entailment Modeling for Translating Unknown Terms. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 791–799, Suntec, Singapore.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69.

Negri, M., Kouylekov, M., and Magnini, B. (2008). Detecting Expected Answer Relations through Textual Entailment. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 532–543, Haifa, Israel.

Nielsen, R. D., Ward, W., and Martin, J. H. (2009). Recognizing Entailment in Intelligent Tutoring Systems. *Natural Language Engineering*, 15:479–501.

Padó, S., Cer, D., Galley, M., Jurafsky, D., and Manning, C. D. (2009). Measuring Machine Translation Quality as Semantic Equivalence: A Metric based on Entailment Features. *Machine Translation*, 23(2–3):181–193.

Padó, S., de Marneffe, M.-C., MacCartney, B., Rafferty, A. N., Yeh, E., and Manning, C. D. (2008). Deciding Entailment and Contradiction with Stochastic and Edit Distance-based Alignment. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Pantel, P., Bhagat, R., Coppola, B., Chklovski, T., and Hovy, E. (2007). ISP: Learning Inferential Selectional Preferences. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 564–571, Rochester, New York, USA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, Philadelphia, Pennsylvania, USA.

Pennacchiotti, M. and Pantel, P. (2009). Entity Extraction via Ensemble Semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 238–247, Suntec, Singapore.

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to Resolve Bridging References. In *Proceedings of ACL*, pages 143–150, Barcelona, Spain.

Romano, L., Kouylekov, M. O., Szpektor, I., Dagan, I., and Lavelli, A. (2006). Investigating a Generic Paraphrase-Based Approach for Relation Extraction. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 409–416, Trento, Italy.

Roth, D. and Sammons, M. (2007). Semantic and Logical Inference Model for Textual Entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 107–112, Prague, Czech Republic.

Schoenmackers, S., Etzioni, O., Weld, D. S., and Davis, J. (2010). Learning First-order Horn clauses from Web Text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1088–1098, Cambridge, Massachusetts, USA.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., and Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Comput.*, 13:1443–1471.

Sekine, S. (2005). Automatic Paraphrase Discovery Based on Context and Keywords between NE Pairs. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP2005)*, Jeju Island, South Korea.

Shnarch, E., Barak, L., and Dagan, I. (2009). Extracting Lexical Reference Rules from Wikipedia. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09, pages 450–458, Suntec, Singapore.

Shnarch, E., Goldberger, J., and Dagan, I. (2011a). A Probabilistic Modeling Framework for Lexical Entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA.

Shnarch, E., Goldberger, J., and Dagan, I. (2011b). Towards a Probabilistic Model for Lexical Entailment. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 10–19, Edinburgh, Scotland, UK.

Snow, R., Jurafsky, D., and Ng, A. Y. (2004). Learning Syntactic Patterns for Automatic Hypernym Discovery. In *Advances in Neural Information Processing Systems (NIPS 2004)*, Granada, Spain.

Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 801–808, Sydney, Australia.

Stokoe, C., Oakes, M. P., and Tait, J. (2003). Word Sense Disambiguation in Information Retrieval Revisited. In *Proceedings of the 26th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval*, pages 159–166, Toronto, Canada.

Surdeanu, M., McClosky, D., Smith, M. R., Gusev, A., and Manning, C. D. (2011). Customizing an Information Extraction System to a New Domain. In *Proceedings of the Workshop on Relational Models of Semantics*, Portland, Oregon, USA.

Szpektor, I., Dagan, I., Bar-Haim, R., and Goldberger, J. (2008). Contextual Preferences. In *Proceedings of ACL-08: HLT*, pages 683–691, Columbus, Ohio, USA.

Szpektor, I., Tanev, H., Dagan, I., and Coppola, B. (2004). Scaling Web-based Acquisition of Entailment Relations. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP*, pages 41–48, Barcelona, Spain.

Wang, R. and Neumann, G. (2008). A divide-and-conquer Strategy for Recognizing Textual Entailment. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, Maryland, USA.

Webber, B., Egg, M., and Kordoni, V. (2010). Discourse Structure: Theory, Practice and Use. ACL 2010 tutorial.

Yates, A. and Etzioni, O. (2009a). Unsupervised Methods for Determining Object and Relation Synonyms on the Web. *Journal of Artificial Intelligence Research (JAIR)*, 34:255–296.

Yates, A. and Etzioni, O. (2009b). Unsupervised Methods for Determining Object and Relation Synonyms on the Web. *Journal of Artificial Intelligence Research (JAIR)*, 34:255–296.

Zanzotto, F. M., Pennacchiotti, M., and Moschitti, A. (2009). A Machine Learning Approach to Textual Entailment Recognition. *Journal of Natural Language Engineering*, 15(4):551–582.

Zhang, M., Lin, C., and Ma, S. (2004). How Effective Is Query Expansion for Finding Novel Information? In *Proceedings of IJCNLP*, pages 149–157, Hainan Island, China.

Zhitomirsky-Geffet, M. and Dagan, I. (2009). Bootstrapping Distributional Feature Vector Quality. *Computational Linguistics*, 35(3):435–461.