UB.dmirg: A Syntactic Lexical System for Recognizing Textual Entailments

Bahadorreza Ofoghi
Centre for Informatics & Applied Optimization, University of Ballarat
Victoria 3350, Australia
b.ofoghi@ballarat.edu.au

John Yearwood
Centre for Informatics & Applied Optimization, University of Ballarat
Victoria 3350, Australia
j.yearwood@ballarat.edu.au

Abstract

This paper reports on our *Recognizing Textual Entailment* (RTE) system developed for participation in the Text Analysis Conference RTE 2009 competition. The development of the system is based on the lexical entailment between two text excerpts, namely the *hypothesis* and the *text*. To extract atomic parts of hypotheses and texts, we carry out syntactic parsing on the sentences. We then utilize WordNet and FrameNet lexical resources for estimating lexical coverage of the text on the hypothesis. Using a failure analysis process, we show that the main difficulty of our RTE system relates to the underlying difficulty of syntactic analysis of sentences.

1 Introduction

Success in many automated natural language applications implies an accurate understanding of the meaning (semantics) of texts underlying the surface structures (syntax) by machines. This becomes challenging with different syntactic forms and dissimilar terms and phrases expressing the same semantics. Automated natural language applications make extensive use of fine-grained text processing modules that enable them in more effective dealings with structurally complicated texts.

One of the current text processing tasks is concerned with inferring the truth or falsity of a piece of text from the evidence that is formulated in another potentially larger text excerpt. This has now become a direction of study for the members of the natural language processing community and is known as *Recognizing Textual Entailment* (RTE). The problem of RTE is formally described as recognizing the relationship between a pair of texts referred to as *hypothesis* and *text*. The hypothesis (H) is a succinct piece of text and the text (T) includes a few sentences the meaning of which may or may not entail the truth of the predicate(s) in the hypothesis.

Different natural language applications may utilize a RTE module in order to relate syntactically different texts that bear similar meanings[1]. For instance, for a Question Answering (QA) system, it becomes an important step to understand that the text "*In 1974, using beams of electrons and antielectrons, or positrons, Richter discovered particle that came to be called Psi/J. It contained two quarks possessing a previously unknown flavor called charm...*" entails the meaning of the answer to the question "*Who discovered Psi/J?*" and then to try to extract the exact answer "*Richter*".

2 Related work

A few approaches to RTE have been developed during recent years. This includes the following.

Term-based approach – Most of the systems that take this approach consider morphological and

[1]Examples of the natural language applications that may embed RTE include Question Answering, Information Extraction, Text Summarization, and Machine Translation.

lexical variations of the terms in texts and hypotheses and determine the existence of entailment between the texts and hypotheses by means of their lexical similarities (Braz et al., 2005; Pazienza et al., 2005; Rodrigo et al., 2008).

Logic-proving approach – The systems that follow this approach apply elements of classical or plausible logic to infer whether the concepts present in the text entail the truth of the hypothesis. The logical procedures are called on a number of feature elements of the texts and hypotheses such as propositions or other logic forms (Akhmatova and Molla, 2006; Tatu and Moldovan, 2005; Clark and Harrison, 2008).

Syntax-based approach – Some existing systems carry out a similarity analysis between the dependency trees extracted from the texts and hypotheses in order to identify the entailment relationships (Lin and Pantel, 2001; Kouylekov and Magnini, 2005; Yatbaz, 2008). There are also systems that take a *paraphrase detection* strategy to generate a set of different styles of the hypotheses with the aim of searching for a subset of which may occur in the texts (Bosma and Callison-Burch, 2006).

Semantic role-based approach – There are systems that annotate the sentences of the texts and hypotheses with semantic roles (using *shallow semantic parsers*) and then analyze the coincidences between sets of assigned semantic roles (Braz et al., 2005).

Knowledge-based approach – The utilization of world knowledge in these systems facilitates recognizing entailment relationships where existing lexical or semantic knowledge is not adequate for confidently inferring the relationships. One available structure that is moving towards formulating world knowledge is Cyc[2]. We have not found any previous RTE system that uses Cyc despite its obvious potential utility in natural language applications (Mahesh et al., 1996).

Generally, all of the system types mentioned above take either a *forward* or a *backward* methodology for identifying the entailment relationships. In the forward methodology, the system begins with analyzing the hypothesis and generating variants or extracting features that are eventually used to analyze the text. Systems that detect paraphrases of the hypothesis sentences fall in this category. In the backward methodology, however, the system analyzes the text features first and relates extracted features to the hypothesis. Most of the systems mentioned earlier follow this methodology.

Our RTE system takes the term-based (lexical) approach to make decisions about textual entailment relationships. And, in contrast with the two methodologies mentioned, it analyzes the hypothesis and the text in parallel. Details of our system are given in section 3.

3 System architecture

To identify entailment relationships between texts and hypotheses, we have developed a term-based approach that analyzes both the texts and hypotheses at the *lexical level* and then produces relations at the *text level*. In performing this process, our system follows the functional steps shown in Figure 1.

3.1 Preprocessing

Before performing any type of analysis/process on the pairs of hypotheses and texts, some preprocessing steps are taken. The preprocessing stage is necessary in order for sentence extraction and the syntactic analysis of the sentences to be successfully carried out. Our RTE system performs the following preprocessing tasks:

- If the hypothesis/text does not finish with a ".", then a "." is added to the end of the hypothesis/text.
- If the hypothesis/text does not start with a capital letter, then its first letter is changed to the capital form.
- Some grammatical issues are resolved. For instance, every occurrence of the string "*, and*" is replaced with "*and*" and the string "*, as well as*" is replaced with "*as well as*".

3.2 Sentence extraction

Given a pair of preprocessed hypothesis and text, we first extract sentences from each. The

[2]http://www.cyc.com/

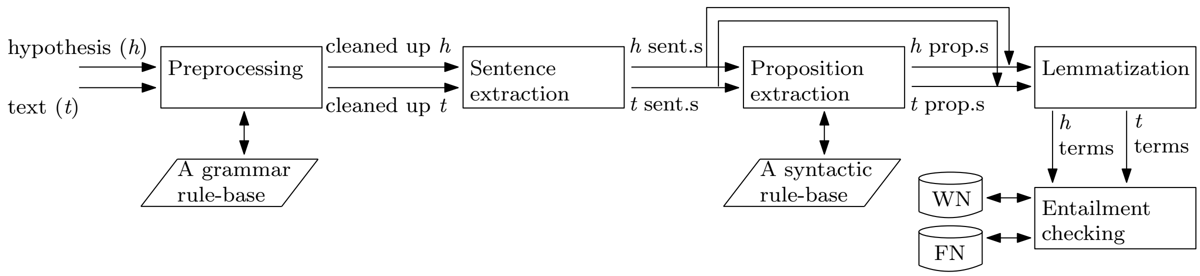

Figure 1: Pipelined functional architecture of our RTE system

hypothesis usually includes only a single sentence; however, the text may contain a few sentences that need to be separated from each other.

We utilize the *LingPipe*[3] sentence splitter which extracts sentences along with their main entity types (locations, people, and organizations). At this stage, our RTE system does not make use of the named entities extracted by LingPipe.

3.3 Proposition extraction

Propositions are extracted from each sentence in the hypothesis and the text. A proposition is an atomic representation of concepts in the texts in which there are no clauses or dependent parts of texts included. For instance, from the sentence "*The girl playing tennis is not my friend.*" the proposition "*girl playing tennis*" can be extracted.

To extract propositions, we use *Link Grammar Parser* (LGP) (Sleator and Temperley, 1993) and follow the procedure explained in (Akhmatova and Molla, 2006). Sentences are parsed using LGP and a set of links between sentence constituents are extracted. The links include syntactic relationships such as subject, object, and prepositional relations. There are seven rules introduced in (Akhmatova and Molla, 2006) and three new rules that we have developed for extracting propositions from sentences. Table 1 shows our new syntactic rules. Given the sentence "*Children are being sexually abused by peacekeepers.*", for instance, the output parse will be like what is shown in Figure 2. From this, we are able to extract the proposition "*peacekeepers abuse children.*". In this particular case, the third rule, that searches for any occurrence of the sequence *Ss/Spx-Pgb-Pv-MVp-Js/Jp* in the output parse by the LGP parser, extracts the proposition.

[3] Alias-i. 2008. LingPipe 3.8.2. http://alias-i.com/lingpipe.

Table 1: Three new syntactic rules for extracting propositions

Linkage	Elements
AN-Mg	AN: connects noun modifiers to nouns, Mg: connects certain prepositions to nouns
AN-Ss/Sp-MVp-Js/Jp	S_: connects subjects to verbs, MVp: connects prepositions to verbs, J_: connects prepositions to their objects
Ss/Spx-Pg*b-Pv-MVp-Js/Jp	Pg*b: connects verbs to present participles, Pv: connects forms of "be" to passive participles

The main advantage of extracting/using propositions is the surface generalization/unification of texts. In other words, both passive and active sentences can, for example, be represented in the same form. This makes the task of semantic alignment of the sentences more effectively achievable by machines.

The other advantage of using propositions is that by extracting brief factual parts of texts into propositions, it is more effectively possible for machines to capture/analyse the integrated meaning of a whole sentence or paragraph.

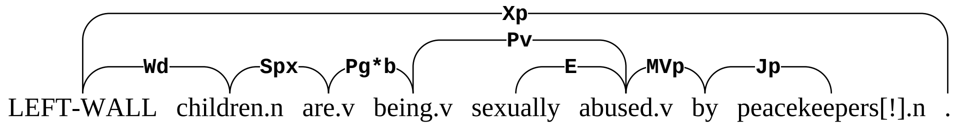

Figure 2: LGP output of the sentence "*Children are being sexually abused by peacekeepers.*"

3.4 Lemmatization

Before semantic alignment is carried out, all hypothesis and text terms are lemmatized using *TreeTagger* (Schmid, 1994). This means that the terms are unified to their single lemma like the transformation of the terms "*abusing*" and "*abused*" to the lemma "*abuse*". The lemmatization step helps the system relate the two propositions "*girl plays tennis*" and "*girl playing tennis*", although they differ in the terms "*plays*" and "*playing*".

3.5 Entailment checking

We finally check the entailment between each pair of propositions extracted from the hypothesis and the text. The idea here is that the truth of each single proposition in the hypothesis needs to be entailed at least by the meaning of a proposition in the text in order for our RTE system to decide whether the meaning of the text entails the truth of the hypothesis.

Checking the pairwise entailment between propositions in our work focuses on the lexical items occurring in the propositions. At this stage, we find the relationships between pairs of lexical items in the propositions regardless of their position. If all lexical items of the hypothesis proposition have related terms in the text proposition, then the decision is that the hypothesis proposition is entailed by the text proposition and an *Entailment* relation is assigned to the pair; otherwise, a *No Entailment* relation is assigned to the hypothesis-text pair.

We use two lexical resources, WordNet (Miller et al., 1990) and FrameNet (Baker et al., 1998), to find relationships between different lexical items. When using WordNet, we assume that a term is semantically interchangeable with its *exact occurrence*, its *synonyms*, and its *hypernyms*. In extracting hypernyms, we only traverse the path in the corresponding WordNet synset for two links. In other words, we exclude the hypernyms that are more distant than two links to the original terms in WordNet synsets.

In utilizing FrameNet, if two lexical items are covered in a single FrameNet frame, then the two items are treated as semantically related in our work. The two verbs "*fly*" and "*pace*", for instance, are covered in (inherited from) the same FrameNet frame "*Self_motion*"; therefore, we assume that these two verbs are semantically interchangeable. This type of event-based similarity is not covered in WordNet.

The frame evocation procedure, in our work, does not take a shallow semantic parsing approach. Instead, a term lookup process finds the set of frames that cover a lexical item. In checking the relatedness of terms using FrameNet, therefore, the exact occurrences of the terms and the set of frames from which the terms inherit are compared against each other. If at least one of these items in a hypothesis proposition have a corresponding item in the text proposition, then the decision is that the item is covered by the text proposition.

In cases where there is no proposition extracted for hypothesis and/or text sentences[4], the whole hypothesis and/or text sentences are taken to the step of entailment checking after their terms are lemmatized. In such cases, we use the *Levenstein* edit Distance (LD) between the hypothesis and the text. We use a shallow procedure where the LD distance takes characters as arguments. If the LD distance between a hypothesis and a text sentence is lower than a pre-defined threshold, then we infer that the text entails the hypothesis. The edit distance,

[4]The LGP syntactic parser may return no parses for sentences that are not grammatical and this leads to the extraction of no propositions in our work.

with different arguments, has previously been used in (Castillo and Alemany, 2008; Pado et al., 2008; Rodrigo et al., 2008).

4 An example scenario

To better understand the architecture of our RTE system, consider the following pair:

text. *His niece was sexually abused by UN peacekeepers and aid workers. Children in post-conflict areas are being abused by the very people drafted into such zones to help look after them, says Save the Children.*

hypothesis. *UN peacekeepers maltreat children.*

Step 1 – In this particular example, both the text and the hypothesis pass the preprocessing step with no changes.

Step 2 – The sentence extraction phase, using LingPipe, returns one sentence for the hypothesis "*UN peacekeepers maltreat children.*" and two sentences for the text "*His niece was sexually abused by UN peacekeepers and aid workers.*" and "*Children in post-conflict areas are being abused by the very people drafted into such zones to help look after them, says Save the Children.*".

Step 3 – Using the LGP parser, from the single sentence of the hypothesis, the proposition "*peacekeepers maltreat children*" is extracted. From the first sentence in the text, the propositions "*peacekeepers abused niece*" and "*workers abused niece*" are extracted. The second sentence in the text cannot be parsed by the LGP parser and therefore, no propositions can be extracted from it.

This step shows the strength of our methodology in terms of unifying different syntactical forms of active and passive sentences. As can be seen, from both the active sentence in the hypothesis and the passive sentence in the text, propositions are extracted that are syntactically uniform and thus comparable (see Step 5).

Step 4 – After lemmatization, propositions look like "*peacekeeper maltreat child*", "*peacekeeper abuse niece*", and "*worker abuse niece*".

Step 5 – The entailment checking phase checks the pairs of propositions of the text and the hypothesis. In this case, three pairs are checked to see whether any of the text propositions entail the hypothesis proposition. This process includes extending proposition terms. The two propositions "*peacekeeper maltreat child*" and "*peacekeeper abuse niece*" differ in the verbs "*maltreat*" and "*abuse*" and in the nouns "*child*" and "*niece*". A synonym lookup process using WordNet and a frame evocation from FrameNet (the frame "*Abusing*") return with "*maltreat*" and "*abuse*" being interchangeable (in our work we call them extensions) of each other. To analyze the relatedness of "*child*" and "*niece*", however, only FrameNet returns with the same semantic frame "*Kinship*" that relates the two lexical items. This is an example of event-based lexical relationship beyond WordNet's synonymy and hyponymy/hypernymy information and can only be achieved using FrameNet.

5 Experiments

5.1 System settings and results

We have developed a *baseline* setting for our RTE system and used it for our experiments prior to the TAC-RTE 2009 competition as well as for the first run of the competition this year. Table 2 shows the baseline system settings and also the changes made for our other runs in the TAC-RTE 2009 competition. In Table 2 *Vs* stands for verbs, *Ns* means nouns, *WNd* represents the WordNet hypernymy distance to be taken while extending lexical items, *LD* is the Levenstein Distance measure used for comparing strings of texts where there is no proposition extracted, and *hypo TC* stands for the term coverage of hypothesis (propositions) that is sought when the lexical-level entailment checking is carried out. To explain one of the further runs, in run3, verbs are extended using FrameNet and WordNet (fn+wn), nouns are also extended using WordNet and FrameNet (wn+fn), the WordNet hypernymy distance is 3, the LD distance is equal to 8 characters, and the entailment checking is carried out on verb and noun phrases only.

Table 2: Settings of our RTE runs, run1 shows the baseline system settings

RTE run	Vs	Ns	WNd	LD	hypo TC
run1	fn	wn	1	3	all
run2	fn+wn	wn+fn	1	3	all
run3	fn+wn	wn+fn	3	8	Vs&Ns

In the TAC-RTE 2008 dataset (rte4_test), there are 1000 pairs of hypotheses and texts in four categories for Question Answering (QA), Information Extraction (IE), and Information Retrieval (IR), and Summarization (SUM) tasks. In the TAC-RTE 2009 datasets (rte5_dev and rte5_test), however, there are only 600 pairs for QA, IE, and IR tasks. We report the *accuracy* of our RTE system for these categories in Table 3. Accuracy for each category cat, denoted by acc_{cat}, is calculated using the formula in Equation 1 where $corr_{cat}$ represents the number of correctly classified pairs in cat and $|cat|$ shows the total number of items in category cat.

$$\begin{aligned} acc_{cat} &= \frac{corr_{cat}}{|cat|} \\ cat &\in \{QA, IE, IR, SUM\} \end{aligned} \tag{1}$$

Table 3: Accuracy of our RTE runs on the RTE4 and RTE5 datasets – Avg. is a macro average

Dataset/run	Accuracy				
	QA	IE	IR	SUM	Avg.
rte4_test/run1	0.480	0.500	0.506	0.490	0.496
rte5_dev/run1	0.480	0.470	0.520	N/A	0.490
rte5_test/run1	0.485	0.505	0.510	N/A	0.500
rte5_test/run2	0.485	0.505	0.510	N/A	0.500
rte5_test/run3	0.485	0.505	0.510	N/A	0.500

More detailed analysis of the results with particular attention to the two classes *Entailment* and *No Entailment* are given in Table 4. In this table, the total performances of the systems are shown. *Recall* for each relationship rel is denoted by rec_{rel} and is calculated using Equation 2 where $corr_{rel}$ is the number of correctly classified pairs into rel and $|rel|$ shows the total number of pairs which should be classified into rel.

$$\begin{aligned} rec_{rel} &= \frac{corr_{rel}}{|rel|} \\ rel &\in \{ent., No\ ent.\} \end{aligned} \tag{2}$$

Table 4: Detailed analysis of our RTE runs on the RTE4 (500 pairs per class) and RTE5 (300 pairs per class) datasets

Dataset/run	Correctly classified		Recall	
	ent.	No ent.	ent.	No ent.
rte4_test/run1	70	426	0.140	0.852
rte5_dev/run1	25	269	0.083	0.896
rte5_test/run1	23	277	0.076	0.923
rte5_test/run2	23	277	0.076	0.923
rte5_test/run3	23	277	0.076	0.923

5.2 Ablation tests

We have submitted ablation runs to the TAC-RTE 2009 competition. Our ablation runs were based on our most complete run, rte5_test/run3 in Table 3, that included both FrameNet and WordNet lexical resources. The three ablation runs were:

- rte5/abl1 = rte5_test/run3-fn: in this run, we excluded FrameNet from the system resources to understand the contribution of FrameNet to the task.
- rte5/abl2 = rte5_test/run3-wn: in this run, we excluded WordNet from the system resources to understand the contribution of WordNet to the task.
- rte5/abl3 = rte5_test/run3-fn-wn: in this run, we excluded both FrameNet and WordNet from the system resources to understand the contribution of these linguistic resources to the task.

Table 5 summarizes the results of our submitted ablation runs.

5.3 Discussion

The results returned from this year's TAC competition (on the RTE5 test dataset) for our RTE

Table 5: Accuracy of our ablation runs on the RTE5 test dataset – Avg. is a macro average

Dataset/run	Accuracy				
	QA	IE	IR	SUM	Avg.
rte5/abl1	0.485	0.505	0.510	N/A	0.500
rte5/abl2	0.485	0.505	0.510	N/A	0.500
rte5/abl3	0.485	0.505	0.510	N/A	0.500

system comes with no surprise as we did not expect an accuracy greater than those of our previous experiments during system training on the RTE4 test and RTE5 development datasets. However, the identical results of our three runs in the TAC competition are surprising.

As shown in Table 3, in our previous runs, our baseline RTE system achieves an average accuracy of 0.496 and 0.490 for the RTE4 test and RTE5 development datasets. An average accuracy of 0.500 on the RTE5 test dataset is, however, our best achievement so far.

A more detailed analysis of these results in Table 4 shows that our RTE system has not been very successful in recognizing correct entailment relationships. On the RTE4 test dataset, the entailment recall of 0.140 for 70 correctly classified items (out of 500 pairs), on the RTE5 development dataset, the entailment recall of 0.083 for only 25 correctly classified items (out of 300 pairs), and on the RTE5 test dataset, the entailment recall of 0.076 for only 23 correctly classified items (out of 300 pairs) do not show high effectiveness in entailment recognition.

It was surprising that our ablation runs have returned with similar results all identical to our baseline run. We never anticipated that the utilization of FrameNet and WordNet will not contribute to the classification performance of our RTE system. After analyzing our system, we found that this was due to a software problem in our system which we have not had time to fix so far.

The overall statistics of the TAC-RTE 2009 for 55 runs submitted by 13 participant teams shows the high, median, and low 2-way classification accuracies of 0.7350, 0.6117, and 0.5000 respectively. The overall performance of our RTE system does not reach high levels of accuracy, compared with the TAC-RTE 2008 participant systems (Giampiccolo et al., 2008) and the TAC-RTE 2009 statistics. We have conducted a failure analysis process to understand the underlying difficulty of the system.

5.4 System failure analysis

We have carried out an error analysis process of our baseline RTE system on the RTE4 test and the RTE5 development and test datasets with respect to the step of syntactic parsing that leads to proposition extraction. Table 6 summarizes the result of this analysis where *hypo* stands for hypothesis and *both* refers to the intersection of the sets of hypotheses and texts. From this table, it can be seen that the major barrier that interferes with our RTE system's performance is the syntactic parsing stage where for the RTE4 test dataset, 131+320-57=394 is the union set of hypotheses and texts for which no parses are returned by the LGP parser. This equates to the fact that the system has access to the parse of only ∼60% of the dataset to extract propositions. For the RTE5 development dataset this ratio is ∼80% of the dataset and for the RTE5 test dataset the ratio is ∼83%.

From another viewpoint, for the RTE4 test dataset, 453+574-261=766 is the total number of hypotheses and texts together where no propositions can be extracted for either the hypothesis or the text sentences. As a result, the semantic expansion process with WordNet and FrameNet and the entailment checking procedure of our baseline RTE system have access to proposition-level information for ∼23% of the pairs in the RTE4 test dataset. For the RTE5 development dataset this ratio is ∼29% of the pairs and for the RTE5 test dataset the ratio is also ∼29%.

The identical results of our three runs in the TAC-RTE 2009 competition and those of the three ablation tests, along with the results of our error analysis suggests that, at this stage, the performance of our RTE method has reached a plateau on the ratio of the dataset for which syntactic parses are returned by the LGP parser and propositions can be extracted.

Table 6: Error analysis of our RTE runs on the RTE4 and RTE5 datasets (only run1)

Dataset	No parse			No prop.		
	hypo	text	both	hypo	text	both
rte4_test	131	320	57	453	574	261
rte5_dev	58	60	2	352	192	119
rte5_test	58	50	2	367	174	116

Therefore, we believe that, to improve the effectiveness of our lexical (term-based) RTE system, there is a need for further elaboration in two aspects:

- *Syntactic parsing*, using a more capable parser that is less sensitive to the grammatical/structural flaws in texts and can more effectively handle long sentences, and
- *Proposition extraction*, by extracting/learning and utilizing a greater number of rules to extract propositions from parsed sentences.

6 Conclusion and future work

A lexical Recognizing Textual Entailment (RTE) system has been introduced in this paper. This 2-way RTE system utilizes a syntactic approach prior to the term-based analysis of the hypotheses and texts in identification of entailment relationships. This syntactic procedure parses sentences extracted from hypotheses and texts and, using a syntactic rule-base, extracts propositions from the sentences. The lexical coverage of the terms in the propositions is checked considering the extensions of terms using WordNet and FrameNet linguistic resources. In cases where no propositions can be extracted from the sentences, the Levenstein Distance is used to estimate entailment relationships.

The results of our RTE system on three datasets of the Text Analysis Conference (TAC) RTE tracks have been reported and shown moderate performances for our system. We have carried out a failure analysis of this RTE system to understand the underlying difficulties that interfere with the system performances. This has shown that the syntactic analysis of the hypotheses and texts, where sentences are parsed and propositions are extracted, is the main challenge that our system faces at this stage.

We have carried out other RTE tests using different settings of our system to obtain a better understanding of how our system may be improved. The results, however, show that no improvements can be achieved by changing minor settings of the system, such as the different combinations of utilization of WordNet and FrameNet for extending different part-of-speech lexical items. This suggests that there is a real need for improving the syntactic analysis stage of our system in order to achieve higher levels of 2-way classification performances by our RTE system.

We are planning to achieve the improvement over the syntactic analysis stage by using a more sophisticated syntactic parser that is less sensitive to grammatical/structural flaws in texts and can better handle long and complicated sentences. We would also like to improve the rule-base that extracts propositions from parsed sentences. After changing the syntactic parser, we may need to re-establish the rule-base considering the features of the new parser.

We are also planning to carry out an in-depth and comprehensive analysis of our methodology of using FrameNet and WordNet in conjunction with proposition-level information to find ways of improving our lexical RTE system.

References

Elena Akhmatova and Diego Molla. 2006. Recognizing textual entailment via atomic propositions. In *Proceedings of the Machine Learning Challenges Workshop (MLCW)*, 385–403. Southampton, UK.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING*, 86–90. Universite de Montreal, Montreal, Quebec, Canada.

W.E. Bosma and C. Callison-Burch. 2006. Paraphrase substitution for recognizing textual entailment. In *Working Notes of CLEF 2006*, 1–8. Alicante, Spain.

Julio Javier Castillo and Laura Alonso i Alemany. 2008. An approach using named entities for recognizing textual entailment. In *Proceedings of the*

Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment. Gaithersburg, Maryland, USA.

Peter Clark and Phil Harrison. 2008. Recognizing textual entailment with logic inference. In *Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment*. Gaithersburg, Maryland, USA.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. 2008. The fourth PASCAL recognizing textual entailment challenge. In *Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment*, 86–90. Gaithersburg, Maryland, USA.

M. Kouylekov and B. Magnini. 2005. Recognizing textual entailment with tree edit distance algorithms. In *Proceedings of the First PASCAL Challenges Workshop on Recognizing Textual Entailment*, 17–20. Southampton, UK.

D. Lin and P. Pantel. 2001. DIRT - Discovery of inference rules from text. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 323–328. San Francisco, California, USA.

K. Mahesh, S. Nirenburg, J. Cowie, and D. Farwell. 1996. An assessment of Cyc for natural language processing. *CRL Technical Report MCCS-96-302*. New Mexico State University.

George A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Sebastian Pado, Marie-Catherine de Marneffe, Bill MacCartney, Anna N. Rafferty, Eric Yeh, and Christopher D. Manning. 2008. Deciding entailment and contradiction with stochastic and edit distance-based alignment. In *Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment*. Gaithersburg, Maryland, USA.

M. T. Pazienza, M. Pennacchiotti, and F. M. Zanzotto. 2005. Textual entailment as syntactic graph distance: A rule based and a SVM based approach. In *Proceedings of the First PASCAL Challenges Workshop on Recognizing Textual Entailment*, 25–28. Southampton, UK.

Alvaro Rodrigo, Anselmo Penas, and Felisa Verdejo. 2008. Towards an entity-based recognition of textual entailment. In *Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment*. Gaithersburg, Maryland, USA.

R. de Salvo Braz, R. Girju, V. Punyakanok, D. Roth, and M. Sammons. 2005. Textual entailment recognition based on dependency analysis and WordNet. In *Proceedings of the First PASCAL Challenges Workshop on Recognizing Textual Entailment*, 29–32. Southampton, UK.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the Conference on New Methods in Language Processing*. Manchester, UK.

Daniel Sleator and Davy Temperley. 1993. Parsing English with a link grammar. In *Proceedings of the Third International Workshop on Parsing Technologies*.

Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, 371–378. Vancouver, British Columbia, Canada.

Mehmet Ali Yatbaz. 2008. RTE4: Normalized dependency tree alignment using unsupervised n-gram word similarity score. In *Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment*. Gaithersburg, Maryland, USA.