From Lexical Entailment to Recognizing Textual Entailment Using Linguistic Resources

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

In this paper, we introduce our *Recognizing Textual Entailment* (RTE) system developed on the basis of 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. We report the results of our RTE runs on the Text Analysis Conference RTE datasets. Using a failure analysis process, we also 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 meaning of a piece of text from that of 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 meaning of the hypothesis. If the meaning of H can be inferred from that of T, then the relationship is denoted by $T \rightarrow H$. For instance, given H="*UN peacekeepers abuse children.*" and T="*Children as young as six are being sexually abused by UN peacekeepers and aid workers, says a leading UK charity.*" the relation $T \rightarrow H$ holds true.

The classification of the relationship between the hypothesis and the text can be either a 3-way classification or a 2-way classification task. The 3-way classes are:

- *Entailment*: where $T \rightarrow H$.
- *Contradiction*: where $T \rightarrow \neg H$.
- *Unknown*: where there is not enough evidence available in the text to decide whether $T \rightarrow H$ or $T \rightarrow \neg H$.

In the 2-way classification method, the *Contradiction* and *Unknown* relations are unified into a single class called *No Entailment*. Our RTE system only considers the 2-way classification task.

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 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 meaning of the text entails that 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[1]. We have not found any previous RTE system that uses Cyc.

Our RTE system takes the term-based (lexical) approach to make decisions about textual entailment relationships.

3 System architecture

3.1 Preprocessing and sentence extraction

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 some basic grammatical and punctuation fixes, such as adding a "." at the end of sentences if the "." is missing or capitalizing the first letter of a sentence if necessary.

We utilize the *LingPipe*[2] sentence splitter to extract sentences from hypotheses and texts.

3.2 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.

Table 1: 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

To extract propositions, we use *Link Grammar Parser* (LGP) (Sleator and Temperley, 1993) and follow the procedure explained in (Akhmatova and Molla, 2006). There are seven rules introduced in (Akhmatova and Molla, 2006) and three new rules that we have developed for extracting propositions. 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 1. From this, we are able to extract the proposition "*peacekeepers abuse children.*".

3.3 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*".

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

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

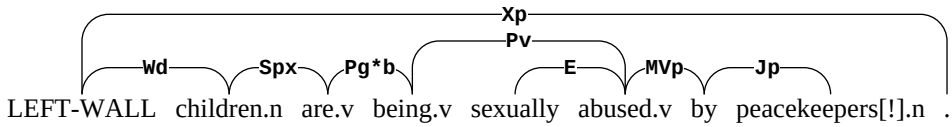

Figure 1: LGP output of the sentence "*Children are being sexually abused by peacekeepers.*"

3.4 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 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 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 encoded in WordNet.

In cases where there is no proposition extracted for hypothesis and/or text sentences, 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 predefined threshold, then we infer that the text entails the hypothesis.

4 Experiments

4.1 Data

We have run our RTE system on three datasets provided by the *Text Analysis Conference* (TAC)[3] for the RTE track:

- TAC-RTE 2008 *test* dataset (rte4 - test), that includes 1000 pairs of hypotheses and texts.
- TAC-RTE 2009 main task *development* dataset (rte5 - dev.), that includes 600 pairs of hypotheses and texts.
- TAC-RTE 2009 main task *test* dataset (rte5 - test), that includes 600 pairs of hypotheses and texts.

4.2 Results

We have carried out experiments with our *baseline* RTE system where:

- The verbs are extended using FrameNet,
- The noun phrases are extended using WordNet,
- The WordNet distance threshold for finding hypernyms is equal to 1,
- The LD distance, in cases where proposition extraction fails, is equal to 3, and
- The term coverage procedure considers all terms in hypotheses (propositions) to have corresponding terms in texts (propositions).

In the TAC-RTE 2008 dataset, there are four categories of hypothesis-text pairs for Question Answering (QA), Information Extraction (IE), and Information Retrieval (IR), and Summarization (SUM)

[3] http://www.nist.gov/tac/

tasks. In the TAC-RTE 2009 datasets, however, there are only pairs for QA, IE, and IR tasks. We report the *accuracy* and the *recall* of our RTE system for these categories and the two classes *Entailment* and *No Entailment* in Table 2 and Table 3. For the RTE5 test dataset, we still do not have access to the answer set; therefore, recall cannot be measured at this stage.

Table 2: Accuracy of our baseline RTE runs on the RTE4 and RTE5 datasets – Avg. is a macro average

Dataset	Accuracy				
	QA	IE	IR	SUM	Avg.
rte4 - test	0.480	0.500	0.506	0.490	0.496
rte5 - dev.	0.480	0.470	0.520	N/A	0.490
rte5 - test	0.485	0.505	0.510	N/A	0.500

Table 3: Detailed analysis of our baseline RTE runs on the RTE4 and RTE5 datasets

Dataset	Correctly classified		Recall	
	ent.	No ent.	ent.	No ent.
rte4 - test	70	426	0.140	0.852
rte5 - dev.	25	269	0.083	0.896
rte5 - test	N/A	N/A	N/A	N/A

4.3 Discussion

As shown in Table 2, an average accuracy of 0.500 on the RTE5 test dataset is our best achievement so far where 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.

A more detailed analysis of these results in Table 3 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) and on the RTE5 development dataset, the entailment recall of 0.083 for only 25 correctly classified items (out of 300 pairs) do not show high effectiveness in entailment recognition. Although with the accuracy measures obtained for the RTE5 test dataset we expect to see comparable classification performance and recall measures for the RTE5 test dataset, we do not have access to the gold standard test set and cannot report on these items for this dataset.

The overall statistics of the TAC-RTE 2009 systems 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 2009 statistics. We have conducted a failure analysis process to understand the underlying difficulty of the system.

4.4 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 particular attention to syntactic parsing leading to proposition extraction. Table 4 summarizes the result of this analysis where *hypo* stands for hypothesis and *both* refers to the intersection of the sets of hypotheses and texts. The major barrier that interferes with our RTE system's performance seems to be the syntactic parsing stage where for the RTE4 test dataset, there are 131+320-57=394 hypotheses and texts for which no parses are returned by LGP. Therefore, the system has access to the parse of only $\sim$60% of the dataset to extract propositions. For the RTE5 development dataset this ratio is $\sim$80% of the dataset.

From another viewpoint, for the RTE4 test dataset there are 453+574-261=766 hypotheses and texts together where no propositions can be extracted for either the hypothesis or the text sentences. As a result, the semantic expansion and entailment checking procedures have access to proposition-level information for $\sim$23% of the pairs in the RTE4 test dataset. For the RTE5 development dataset, this ratio is $\sim$29% of the pairs.

Table 4: Error analysis of our baseline RTE runs on the RTE4 test and the RTE5 development datasets

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

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 grammati-

cal/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.

5 Conclusion

A lexical Recognizing Textual Entailment (RTE) system participated in the Text Analysis Conference (TAC) 2009 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.

The results of our RTE system on three datasets of the 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.

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.

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.

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.

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.

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.

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

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

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.