WHU at TAC 2009: A Tri-categorization Approach to Textual Entailment Recognition

Han Ren, Donghong Ji
School of Computer Science
Wuhan University, Wuhan 430079, China
cslotus@mail.whu.edu.cn
donghong_ji@yahoo.com

Jing Wan
Center for Study of Language & Information
Wuhan University, Wuhan 430079, China
Jennifer_wanj@yahoo.com.cn

Abstract

This paper describes our system of recognizing textual entailment for RTE-5 challenge at TAC 2009. We propose a textual entailment recognition framework and implement a system of classification which takes lexical, syntactic and semantic features as considered. To improve the performance, some lexical-semantic resources and web knowledge bases are also incorporated in the system. Official results show that our system achieves a medium performance of all participating systems.

1 Indroduction

Given a text fragment, the goal of Textual Entailment is to recognize a hypothesis that can be inferred from it or not. Textual Entailment is a notable field of research that are leveraged in many natural language processing areas, such as document summarization, information retrieval and question answering (Harabagiu and Hickl, 2006).

Following the previous RTE challenge (Giampiccolo *et al.*, 2008), RTE-5 identically adopts a two-way determination and a three-way one, except the ablation tests, which are introduced to evaluate the contribution of each resource to participants' system performances. We participate in the three-way evaluation and the responses are also evaluated by two-way determination automatically.

We propose a textual entailment recognition framework and implement a system for RTE-5 challenge. The system is based on a Maximum Entropy classifier, considering that recognizing entailed relations (Entailment, Contradiction and Unknown) can be viewed as a tri-categorization problem, and makes use of linguistic-based and statistical-based features, and some hybrid features for classification. To improve the performance, some lexical-semantic resources and web knowledge bases are also incorporated in the system. Although large resources and background knowledge such as paraphrase collection and geographic ontology contribute to the better performance (Hickl *et al.*, 2007; Shen *et al.*, 2008), we still employ some basic resources, by which a comparable baseline system can be achieved.

The rest of the paper is organized as follows. In Section 2 the architecture and each part of the system are described. Section 3 gives the experimental results and some discussions. Finally, some conclusions and future work are given.

2 System Overview

The overall architecture of the system is shown in Figure 1, which contains a preprocessing module, a feature extraction module and a classifier. For feature extraction, the training and testing modules utilize the same methods. Procedures of the system is described as follows:

1) for each text fragment and hypothesis, a preprocessing is performed, including stemming, part-of-speech tagging, named entity recognition and anaphora resolution;

2) for feature extraction, lexical, syntactic and semantic features are selected. Additionally, named entity relations are employed as specific features to identify explicit entailed relations;

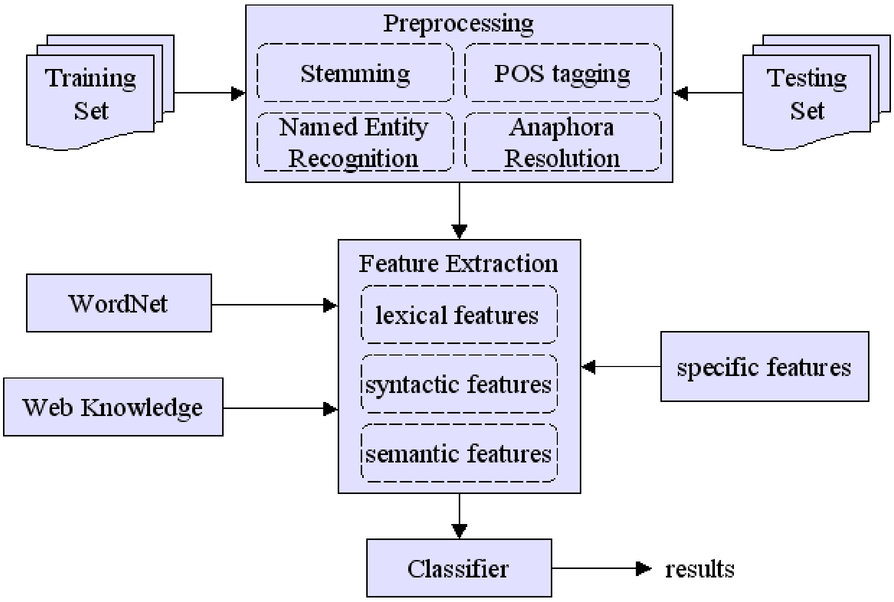

Figure 1. System Architecture

3) all features are provided to the classification module; after that, an entailed relation is given for the text fragment and the hypothesis.

2.1 Preprocessing

We make use of several free tools to deal with the preprocessing:

1)Stemming. Each pair is stemmed by Porter algorithm. Here we use a java version provided by M.F Porter[1].

2)POS Tagging and Named Entity Recognition. We use two state-of-the-art tools, POS Tagger and NE Recognizer that are produced by Stanford NLP group[2], to perform each pair. For named entities, we only consider PERSON, LOCATION and ORGANIZATION.

3)Anaphora Resolution. JavaRap[3] and GuiTAR[4] are combined in our system for anaphora resolution since each one of them achieves a bias result with other modules.

2.2 Lexical Features

For lexical features we consider the following features:

• Bag-Of-Words. The assumption is that *T* and *H* have a closer relationship if they have more identical words.

• *Jaccard* Similarity. Since *cosine* similarity considers the length of a text whereas *T* is always much longer than *H*, it's more appropriate to use *Jaccard* coefficient to compute the similarity of *T* and *H*.

• Longest Common Substring. LCS seems an effective feature to compare the similarity of *T* and *H*, and described as follows:

$$LCS(T,H) = \frac{max\{SubLen(T,H)\}}{min\{Len(T),\ Len(H)\}} \quad (1)$$

• Levenshtein distance. The distance is the minimum number of operations when transforming one string to another. Here we consider tokens instead of characters.

2.3 Syntactic Features

Syntactic features are fetched by mapping *T* and *H* to syntactic parse trees. Since shallow syntactic parsing is more flexible and precise, we use dependency parsing in the system to deal with it. MaltParser[5] is a state-of-the-art dependency parser that yields good results in CoNLL shared tasks. In

[1] http://tartarus.org/~martion/PorterStemmer/

[2] http://nlp.stanford.edu/software/

[3] http://wing.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.html

[4] http://cswww.essex.ac.uk/Research/nle/GuiTAR/

[5] http://maltparser.org/

the system, we use it for syntactic parsing and the training data is derived from syntactic resources such as TreeBank and PropBank. In (Nielsen *et al.*, 2006), the author investigated dependent features which we used part of them in the system: Bag-of-Dependencies, Descendent Relation Features, Combined Verb Descendent Relations, Combined Subject Descendent Relations, Combined Subject-to-Verb Relations and Object Relations.

2.4 Semantic Features

For semantic features we consider lexical-semantic similarity and the predicate-argument structure similarity. We compare all predicate-argument structures in *H* and each sentence in *T* and get the maximum number as the value of the pair's similarity. WordNet is also used to compute the similarity of predicate or arguments and their hyponym. Also, we compute lexical-semantic similarity of *T* and *H* using a method proposed by (Wu *et al.*, 1994).

2.5 Specific Features

For a better performance, named entity relations extracting from web knowledge such as Wikipedia are formed as specific features. The idea is based on (Iftene *et al.*, 2008). For a specified named entities, the author proposed an approach to extract from Wikipedia snippets with named entities related to it.

3 Experimental Results

In RTE-5 challenge, the ratio of three categories are: 0.5 for entailment, 0.15 for contradiction and 0.35 for unknown. For the participation of the challenge, we submitted two runs for three-way evaluation, which differ in the different ratio of lexical, syntactic and semantic similarity. Due to the slight adjustment, results of two runs are almost same, except that the result of 3-way for IE is 0.51. Table 1 shows the results of run 1.

In three-way classification, the overall precision drops 11.16% against two-way classification, while in IR task, the precision drops up to 13.5%. It results from two reasons: 1) the system has a weak discrimination at contradiction pairs, because most of features focus on similarity, while contradiction can also be viewed as a kind of 'similarity' except for negative words. Therefore, some contradiction pairs are identified as entailment pairs rather than contradiction ones; 2) text derived from QA and IE are more precise than IR, since more linguistic judgment are introduced to them. For a better performance, recognizing negation is an important part in the system.

	2-way	3-way
QA	0.62	0.51
IE	0.605	0.515
IR	0.675	0.54
All	0.6333	0.5217

Table 1. Official results of submission for 2-way and 3-way

For evaluating the contribution of each resource to participants' system performances, ablation tests is required by organizer. For this purpose, we revise our system by removing features motivated by PropBank and Wikipedia, and other modules are based on run 1. Table 2 and Table 3 show the results.

	2-way	3-way
QA	0.6	0.48
IE	0.585	0.48
IR	0.655	0.51
All	0.6133	0.49

Table 2. Ablation test removing PropBank

	2-way	3-way
QA	0.615	0.485
IE	0.585	0.48
IR	0.66	0.5
All	0.62	0.4883

Table 3. Ablation test removing Wikipedia

From Table 2 we can see that the overall performance of three-way evaluation drops 2% when removing PropBank. It indicates that PropBank mainly contribute to the learning of syntactic parser while it has not a distinct contribution to textual entailment recognition. An entailment system acquires good syntactic features regardless the training data the syntactic parser employed. In Table 3, the overall performance drops 3.84% when removing Wikipedia. It makes clear that named entities relation can be a help to improve entailment recognition. This conclusion can be proved by many systems in RTE challenges.

4 Conclusion

In this paper, we described our system for RTE-5 challenge at TAC 2009. We propose a textual entailment recognition framework and implement a system of classification which takes lexical, syntactic and semantic features as considered. Official results show that our system achieves a medium performance of all participating systems.

In the future work, recognizing negation is a direction for improving our system since the performance of three-way evaluation is much lower than that of two-way evaluation. On the other hand, recognizing named entities relations can also be improved to increase the precision of the system. For this purpose, some available resources such as Polarity Lexicons and paraphrase collections can be used in the system for a better performance.

Acknowledgments

This research is supported by Natural Science Foundation of China(Grant Nos. 60773011, 90820005) and Self-research program of Doctoral Candidates(including Mphil-PhD) of Wuhan University in 2008.

References

Sanda Harabagiu and Andrew Hickl. 2006. Methods for Using Textual Entailment in Open-Domain Question Answering. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, July 2006, pp. 905-912.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini and Ido Dagan. 2008. The Fourth PASCAL Recognizing Textual Entailment Challenge. In Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment.

Andrew Hickl and Jeremy Bensley. 2007. A Discourse Commitment-Based Framework for Recognizing Textual Entailment. In Proceedings of the Third PASCAL Challenges Workshop on Recognizing Textual Entailment.

Rongzhou Shen, Thade Nahnsen, Claire Grover and Ewan Klein. 2008. Recognising Textual Entailment Focusing on Non-Entailing Text and Hypothesis. In Proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment.

Rodney D. Nielsen, Wayne Ward and James H. Martin. 2006. Toward Dependency Path based Entailment. In Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment.

Zhibiao Wu and Martha Palmer. 1994. Verb Semantics and Lexical Selection. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. Las Cruces, New Mexico.

Adrian Iftene and Alexandra Balahur-Dobrescu. 2008. Named Entity Relation Mining Using Wikipedia. In Proceedings of the Sixth Language Resources and Evaluation Conference, Marrakech, Morocco, 2008.