Modeling Human Inference Process for Textual Entailment Recognition

Hen-Hsen Huang **Kai-Chun Chang** **Hsin-Hsi Chen**
Department of Computer Science and Information Engineering
National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan
{hhhuang, kcchang}@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

Abstract

This paper aims at understanding what human think in textual entailment (*TE*) recognition process and modeling their thinking process to deal with this problem. We first analyze a labeled RTE-5 test set and find that the negative entailment phenomena are very effective features for *TE* recognition. Then, a method is proposed to extract this kind of phenomena from text-hypothesis pairs automatically. We evaluate the performance of using the negative entailment phenomena on both the English RTE-5 dataset and Chinese NTCIR-9 RITE dataset, and conclude the same findings.

1 Introduction

Textual Entailment (*TE*) is a directional relationship between pairs of text expressions, text (*T*) and hypothesis (*H*). If human would agree that the meaning of *H* can be inferred from the meaning of *T*, we say that *T* entails *H* (Dagan et al., 2006). The researches on textual entailment have attracted much attention in recent years due to its potential applications (Androutsopoulos and Malakasiotis, 2010). Recognizing Textual Entailment (*RTE*) (Bentivogli, et al., 2011), a series of evaluations on the developments of English *TE* recognition technologies, have been held seven times up to 2011. In the meanwhile, *TE* recognition technologies in other languages are also underway (Shima, et al., 2011).

Sammons, et al., (2010) propose an evaluation metric to examine the characteristics of a *TE* recognition system. They annotate text-hypothesis pairs selected from the RTE-5 test set with a series of linguistic phenomena required in the human inference process. The *RTE* systems are evaluated by the new indicators, such as how many *T-H* pairs annotated with a particular phenomenon can be correctly recognized. The indicators can tell developers which systems are better to deal with *T-H* pairs with the appearance of which phenomenon. That would give developers a direction to enhance their *RTE* systems.

Such linguistic phenomena are thought as important in the human inference process by annotators. In this paper, we use this valuable resource from a different aspect. We aim at knowing the ultimate performance of *TE* recognition systems which embody human knowledge in the inference process. The experiments show five negative entailment phenomena are strong features for *TE* recognition, and this finding confirms the previous study of Vanderwende et al. (2006). We propose a method to acquire the linguistic phenomena automatically and use them in *TE* recognition.

This paper is organized as follows. In Section 2, we introduce linguistic phenomena used by annotators in the inference process and point out five significant negative entailment phenomena. Section 3 proposes a method to extract them from *T-H* pairs automatically, and discuss their effects on *TE* recognition. In Section 4, we extend the methodology to the BC (binary class subtask) dataset distributed by NTCIR-9 RITE task (Shima, et al., 2011) and discuss their effects on *TE* recognition in Chinese. Section 5 concludes the remarks.

2 Human Inference Process in TE

We regard the human annotated phenomena as features in recognizing the binary entailment relation between the given *T-H* pairs, i.e., ENTAILMENT and NO ENTAILMENT. Total 210 *T-H* pairs are chosen from the RTE-5 test set by Sammons et al. (2010), and total 39 linguistic phenomena divided into the 5 aspects, including knowledge domains, hypothesis structures, inference phenomena, negative entailment phenome-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 446–450,
Sofia, Bulgaria, August 4-9 2013.

na, and knowledge resources, are annotated on the selected dataset.

2.1 Five aspects as features

We train SVM classifiers to evaluate the performances of the five aspects of phenomena as features for *TE* recognition. LIBSVM RBF kernel (Chang and Lin, 2011) is adopted to develop classifiers with the parameters tuned by grid search. The experiments are done with 10-fold cross validation.

For the dataset of Sammons et al. (2010), two annotators are involved in labeling the above 39 linguistic phenomena on the *T-H* pairs. They may agree or disagree in the annotation. In the experiments, we consider the effects of their agreement. Table 1 shows the results. Five aspects are first regarded as individual features, and are then merged together. Schemes "Annotator A" and "Annotator B" mean the phenomena labelled by annotator A and annotator B are used as features respectively. The "A AND B" scheme, a strict criterion, denotes a phenomenon exists in a *T-H* pair only if both annotators agree with its appearance. In contrast, the "A OR B" scheme, a looser criterion, denotes a phenomenon exists in a *T-H* pair if at least one annotator marks its appearance.

We can see that the aspect of *negative entailment phenomena* is the most significant feature among the five aspects. With only 9 phenomena in this aspect, the SVM classifier achieves accuracy above 90% no matter which labeling schemes are adopted. Comparatively, the best accuracy in RTE-5 task is 73.5% (Iftene and Moruz, 2009). In negative entailment phenomena aspect, the "A OR B" scheme achieves the best accuracy. In the following experiments, we adopt this labeling scheme.

2.2 Negative entailment phenomena

There is a large gap between using negative entailment phenomena and using the second effective features (i.e., inference phenomena). Moreover, using the negative entailment phenomena as features only is even better than using all the 39 linguistic phenomena. We further analyze which negative entailment phenomena are more significant.

There are nine linguistic phenomena in the aspect of negative entailment. We take each phenomenon as a single feature to do the two-way textual entailment recognition. The "A OR B" scheme is applied. Table 2 shows the experimental results.

	Annotator A	Annotator B	A AND B	A OR B
Knowledge Domains	50.95%	52.38%	52.38%	50.95%
Hypothesis Structures	50.95%	51.90%	50.95%	51.90%
Inference Phenomena	74.29%	72.38%	72.86%	74.76%
Negative Entailment Phenomena	97.14%	95.71%	92.38%	97.62%
Knowledge Resources	69.05%	69.52%	67.62%	69.52%
ALL	97.14%	92.20%	90.48%	97.14%

Table 1: Accuracy of recognizing binary TE relation with the five aspects as features.

Phenomenon ID	Negative entailment Phenomenon	Accuracy
0	Named Entity mismatch	60.95%
1	Numeric Quantity mismatch	54.76%
2	Disconnected argument	55.24%
3	Disconnected relation	57.62%
4	Exclusive argument	61.90%
5	Exclusive relation	56.67%
6	Missing modifier	56.19%
7	Missing argument	69.52%
8	Missing relation	68.57%

Table 2: Accuracy of recognizing TE relation with individual negative entailment phenomena.

The 1st column is phenomenon ID, the 2nd column is the phenomenon, and the 3rd column is the accuracy of using the phenomenon in the binary classification. Comparing with the best accuracy 97.62% shown in Table 1, the highest accuracy in Table 2 is 69.52%, when missing argument is adopted. It shows that each phenomenon is suitable for some *T-H* pairs, and merging all negative entailment phenomena together achieves the best performance.

We consider all possible combinations of these 9 negative entailment phenomena, i.e., C_1^9+…+ C_9^9 =511 feature settings, and use each feature setting to do 2-way entailment relation recognition by LIBSVM. The notation C_n^m denotes a set of $\frac{m!}{(m-n)!n!}$ feature settings, each with *n* features.

The model using all nine phenomena achieves the best accuracy of 97.62%. Examining the combination sets, we find phenomena IDs 3, 4, 5, 7 and 8 appear quite often in the top 4 feature settings of each combination set. In fact, this setting achieves an accuracy of 95.24%, which is the best performance in C_5^9 combination set. On the one hand, adding more phenomena into (3, 4, 5, 7, 8) setting does not have much performance difference.

In the above experiments, we do all the analyses on the corpus annotated with linguistic phenomena by human. We aim at knowing the ulti-

mate performance of *TE* recognition systems embodying human knowledge in the inference. The human knowledge in the inference cannot be captured by *TE* recognition systems fully correctly. In the later experiments, we explore the five critical features, (3, 4, 5, 7, 8), and examine how the performance is affected if they are extracted automatically.

3 Negative Entailment Phenomena Extraction

The experimental results in Section 2.2 show that disconnected relation, exclusive argument, exclusive relation, missing argument, and missing relation are significant. We follow the definitions of Sammons et al. (2010) and show them as follows.

(a) Disconnected Relation. The arguments and the relations in Hypothesis (*H*) are all matched by counterparts in Text (*T*). None of the arguments in *T* is connected to the matching relation.

(b) Exclusive Argument. There is a relation common to both the hypothesis and the text, but one argument is matched in a way that makes *H* contradict *T*.

(c) Exclusive Relation. There are two or more arguments in the hypothesis that are also related in the text, but by a relation that means *H* contradicts *T*.

(d) Missing Argument. Entailment fails because an argument in the Hypothesis is not present in the Text, either explicitly or implicitly.

(e) Missing Relation. Entailment fails because a relation in the Hypothesis is not present in the Text, either explicitly or implicitly.

To model the annotator's inference process, we must first determine the arguments and the relations existing in *T* and *H*, and then align the arguments and relations in *H* to the related ones in *T*. It is easy for human to find the important parts in a text description in the inference process, but it is challenging for a machine to determine what words are important and what are not, and to detect the boundary of arguments and relations. Moreover, two arguments (relations) of strong semantic relatedness are not always literally identical.

In the following, a method is proposed to extract the phenomena from *T-H* pairs automatically. Before extraction, the English *T-H* pairs are pre-processed by numerical character transformation, POS tagging, and dependency parsing with Stanford Parser (Marneffe, et al., 2006; Levy and Manning, 2003), and stemming with NLTK (Bird, 2006).

3.1 A feature extraction method

Given a *T-H* pair, we first extract 4 sets of noun phrases based on their POS tags, including {noun in *H*}, {named entity (nnp) in *H*}, {compound noun (cnn) in *T*}, and {compound noun (cnn) in *H*}. Then, we extract 2 sets of relations, including {relation in *H*} and {relation in *T*}, where each relation in the sets is in a form of *Predicate*(*Argument*1, *Argument*2). Some typical examples of relations are *verb*(*subject*, *object*) for verb phrases, *neg*(*A*, *B*) for negations, *num*(*Noun*, *number*) for numeric modifier, and *tmod*(*C*, *temporal argument*) for temporal modifier. A predicate has only 2 arguments in this representation. Thus, a di-transitive verb is in terms of two relations.

Instead of measuring the relatedness of *T-H* pairs by comparing *T* and *H* on the predicate-argument structure (Wang and Zhang, 2009), our method tries to find the five negative entailment phenomena based on the similar representation. Each of the five negative entailment phenomena is extracted as follows according to their definitions. To reduce the error propagation which may be arisen from the parsing errors, we directly match those nouns and named entities appearing in *H* to the text *T*. Furthermore, we introduce WordNet to align arguments in *H* to *T*.

(a) Disconnected Relation. If (1) for each $a \in$ {noun in *H*}$\cup${nnp in *H*}$\cup${cnn in *H*}, we can find $a \in T$ too, and (2) for each $r_1=h(a_1,a_2) \in$ {relation in *H*}, we can find a relation $r_2=h(a_3,a_4) \in$ {relation in *T*} with the same header *h*, but with different arguments, i.e., $a_3 \neq a_1$ and $a_4 \neq a_2$, then we say the *T-H* pair has the "Disconnected Relation" phenomenon.

(b) Exclusive Argument. If there exist a relation $r_1=h(a_1,a_2)\in$ {relation in *H*}, and a relation $r_2=h(a_3,a_4)\in$ {relation in *T*} where both relations have the same header *h*, but either the pair (a_1,a_3) or the pair (a_2,a_4) is an antonym by looking up WordNet, then we say the *T-H* pair has the "Exclusive Argument" phenomenon.

(c) Exclusive Relation. If there exist a relation $r_1=h_1(a_1,a_2)\in$ {relation in *T*}, and a relation $r_2=h_2(a_1,a_2)\in$ {relation in *H*} where both relations have the same arguments, but h_1 and h_2 have the opposite meanings by consulting WordNet, then we say that the *T-H* pair has the "Exclusive Relation" phenomenon.

(d) Missing Argument. For each argument $a_1 \in$ {noun in H}$\cup$ {nnp in H}$\cup$ {cnn in H}, if there does not exist an argument $a_2 \in T$ such that $a_1=a_2$, then we say that the *T-H* pair has "Missing Argument" phenomenon.

(e) Missing Relation. For each relation $r_1=h_1(a_1,a_2) \in$ {relation in H}, if there does not exist a relation $r_2=h_2(a_3,a_4) \in$ {relation in T} such that $h_1=h_2$, then we say that the *T-H* pair has "Missing Relation" phenomenon.

3.2 Experiments and discussion

The following two datasets are used in English TE recognition experiments.

(a) 210 pairs from part of RTE-5 test set. The 210 *T-H* pairs are annotated with the linguistic phenomena by human annotators. They are selected from the 600 pairs in RTE-5 test set, including 51% ENTAILMENT and 49% NO ENTAILMENT.

(b) 600 pairs of RTE-5 test set. The original RTE-5 test set, including 50% ENTAILMENT and 50% NO ENTAILMENT.

Table 3 shows the performances of *TE* recognition. The "Machine-annotated" and the "Human-annotated" columns denote that the phenomena annotated by machine and human are used in the evaluation respectively. Using "Human-annotated" phenomena can be seen as the upper-bound of the experiments. The performance of using machine-annotated features in 210-pair and 600-pair datasets is 52.38% and 59.17% respectively.

Though the performance of using the phenomena extracted automatically by machine is not comparable to that of using the human annotated ones, the accuracy achieved by using only 5 features (59.17%) is just a little lower than the average accuracy of all runs in RTE-5 formal runs (60.36%) (Bentivogli, et al., 2009). It shows that the significant phenomena are really effective in dealing with entailment recognition. If we can improve the performance of the automatic phenomena extraction, it may make a great progress on the textual entailment.

Phenomena	210 pairs		600 pairs
	Machine-annotated	Human-annotated	Machine-annotated
Disconnected Relation	50.95%	57.62%	54.17%
Exclusive Argument	50.95%	61.90%	55.67%
Exclusive Relation	50.95%	56.67%	51.33%
Missing Argument	53.81%	69.52%	56.17%
Missing Relation	50.95%	68.57%	52.83%
All	52.38%	95.24%	59.17%

Table 3: Accuracy of textual entailment recognition using the extracted phenomena as features.

4 Negative Entailment Phenomena in Chinese RITE Dataset

To make sure if negative entailment phenomena exist in other languages, we apply the methodologies in Sections 2 and 3 to the *RITE* dataset in NTCIR-9. We annotate all the 9 negative entailment phenomena on Chinese *T-H* pairs according to the definitions by Sammons et al. (2010) and analyze the effects of various combinations of the phenomena on the new annotated Chinese data. The accuracy of using all the 9 phenomena as features (i.e., C_9^9 setting) is 91.11%. It shows the same tendency as the analyses on English data. The significant negative entailment phenomena on Chinese data, i.e., (3, 4, 5, 7, 8), are also identical to those on English data. The model using only 5 phenomena achieves an accuracy of 90.78%, which is very close to the performance using all phenomena.

We also classify the entailment relation using the phenomena extracted automatically by the similar method shown in Section 3.1, and get a similar result. The accuracy achieved by using the five automatically extracted phenomena as features is 57.11%, and the average accuracy of all runs in NTCIR-9 RITE task is 59.36% (Shima, et al., 2011). Compared to the other methods using a lot of features, only a small number of binary features are used in our method. Those observations establish what we can call a useful baseline for *TE* recognition.

5 Conclusion

In this paper we conclude that the negative entailment phenomena have a great effect in dealing with *TE* recognition. Systems with human annotated knowledge achieve very good performance. Experimental results show that not only can it be applied to the English *TE* problem, but also has the similar effect on the Chinese *TE* recognition. Though the automatic extraction of the negative entailment phenomena still needs a lot of efforts, it gives us a new direction to deal with the *TE* problem.

The fundamental issues such as determining the boundary of the arguments and the relations, finding the implicit arguments and relations, verifying the antonyms of arguments and relations, and determining their alignments need to be further examined to extract correct negative entailment phenomena. Besides, learning-based approaches to extract phenomena and multi-class *TE* recognition will be explored in the future.

Acknowledgments

This research was partially supported by Excellent Research Projects of National Taiwan University under contract 102R890858 and 2012 Google Research Award.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research*, 38:135-187.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2011. The seventh PASCAL recognizing textual entailment challenge. In *Proceedings of the 2011 Text Analysis Conference* (*TAC 2011*), Gaithersburg, Maryland, USA..

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the 2009 Text Analysis Conference* (*TAC 2009*), Gaithersburg, Maryland, USA.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (*COLING-ACL 2006*), pages 69-72.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1-27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. *Lecture Notes in Computer Science*, 3944:177-190.

Adrian Iftene and Mihai Alex Moruz. 2009. UAIC Participation at RTE5. In *Proceedings of the 2009 Text Analysis Conference* (*TAC 2009*), Gaithersburg, Maryland, USA.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (*ACL 2003*), pages 439-446.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *The Fifth International Conference on Language Resources and Evaluation* (*LREC 2006*), pages 449-454.

Mark Sammons, V.G.Vinod Vydiswaran, and Dan Roth. 2010. Ask not what textual entailment can do for you... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (*ACL 2010*), pages 1199-1208, Uppsala, Sweden.

Hideki Shima, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi, and Koichi Takeda. 2011. Overview of NTCIR-9 RITE: Recognizing inference in text. In *Proceedings of the NTCIR-9 Workshop Meeting*, Tokyo, Japan.

Lucy Vanderwende, Arul Menezes, and Rion Snow. 2006. Microsoft Research at RTE-2: Syntactic Contributions in the Entailment Task: an implementation. In *Proceedings of the Second PASCAL Challenges Workshop.*

Rui Wang and Yi Zhang. 2009. Recognizing Textual Relatedness with Predicate-Argument Structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 784–792, Singapore.