Computational Linguistics and Chinese Language Processing
Vol. 19, No. 3, September 2014, pp. 39-54

Modeling Human Inference Process for Textual Entailment Recognition

Hen-Hsen Huang*, Kai-Chun Chang* and Hsin-Hsi Chen*

Abstract

To prepare an evaluation dataset for textual entailment (TE) recognition, human annotators label rich linguistic phenomena on text and hypothesis expressions. These phenomena illustrate implicit human inference process to determine the relations of given text-hypothesis pairs. This paper aims at understanding what human think in TE recognition process and modeling their thinking process to deal with this problem. At first, we analyze a labelled RTE-5 test set which has been annotated with 39 linguistic phenomena of 5 aspects by Mark Sammons *et al*., and find that the negative entailment phenomena are very effective features for TE recognition. Then, a rule-based method and a machine learning method are proposed to extract this kind of phenomena from text-hypothesis pairs automatically. Though the systems with the machine-extracted knowledge cannot be comparable to the systems with human-labelled knowledge, they provide a new direction to think TE problems. We further annotate the negative entailment phenomena on Chinese text-hypothesis pairs in NTCIR-9 RITE-1 task, and conclude the same findings as that on the English RTE-5 datasets.

Keywords: Textual Entailment Recognition, Chinese Processing, Semantic.

1. Introduction

Textual Entailment (TE) is a directional relationship between pairs of text expressions, text (*T*) and hypothesis (*H*). Given a text pair *T* and *H*, if human would consider that the meaning of *H* is right by using the information of *T*, then we can infer *H* from *T* and say that *T* entails *H* (Dagan, Glickman, & Magnini, 2006). (S1) shows an example where *T* entails *H*.

* Department of Computer Science and Information Engineering, National Taiwan University
E-mail: {hhhuang, kcchang}@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

(S1) **T**: Norway's most famous painting, 'The Scream' by Edvard Munch, was recovered Saturday, almost three months after it was stolen from an Oslo museum.
H: Edvard Munch painted 'The Scream'.

Because such an inference is important in many applications (Androutsopoulos & Malakasiotis, 2010), the researches on textual entailment have attracted much attention in recent years. 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. The 9th NTCIR Workshop Meeting first introduced a TE task in Chinese and in Japanese called Recognizing Inference in Text (RITE-1) into the IR series evaluation (Shima *et al*., 2011).

The overall accuracy is used as the only evaluation metric in most TE recognition tasks (Androutsopoulos & Malakasiotis, 2010). However, it is hard to examine the characteristics of a system when only considering its performance by accuracy. Sammons *et al*., (2010) proposed an evaluation metric to examine the characteristics of a TE recognition system. They annotated text-hypothesis pairs selected from the RTE-5 test set with a series of linguistic phenomena required in the human inference process. When annotators assume that some linguistic phenomena appear in their inference process to determine whether T entails H, they would label the T-H pair with these phenomena. 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. On the other hand, that would give developers a direction to enhance RTE systems.

For example, (S2) is an instance that matches the linguistic phenomena Exclusive Relation, and this phenomenon suggests *T* does not entail *H*. More than one argument of *H*, i.e., Venus Williams, Marion Bartoli, 2007, and Wimbledon Championships, appear in *T*, but the relation defeated in *H* contracts the relation triumphed in *T*.

(S2) **T**: Venus Williams triumphed over Marion Bartoli of France 6-4, 6-1 yesterday to win the Women's Singles event at the 2007 Wimbledon Championships. For the first time, an American and Frenchwoman were matched up to compete for the British women's singles title. A Wimbledon champion in 2000, 2001 and 2005, Williams was not the favorite to win the title again this year. Currently ranked 23rd in the world, she entered the tournament in the shadow of her sister, Serena Williams.

H: Venus Williams was defeated by Marion Bartoli at the 2007 Wimbledon Championships.

Such linguistic phenomena are thought as crucial in the human inference process by annotators. In the RITE-2 in the 10th NTCIR Workshop Meeting, some linguistic phenomena for TE in Japanese are reported in the unit task subtask (Watanabe *et al.*, 2013). In a similar manner, types of some linguistic phenomena in Chinese are consulted in the RITE-VAL task in the 11th NTCIR Workshop Meeting[1]. In this paper, we use this valuable resource from a different aspect. Instead of using the labelled linguistic phenomena in the evaluation of TE recognition, 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 may be strong features for TE recognition, and this finding confirms the previous study of Vanderwende *et al* (2006). Moreover, we propose a method to acquire the linguistic phenomena automatically and use them in TE recognition. Our method is evaluated on both the English RTE-5 dataset and the Chinese NTCIR-9 RITE-1 dataset. Experimental results show that our method achieves decent performances near the average performances of RTE-5 and NTCIR-9 RITE-1. Compared to the other methods incorporating a lot of features, only a tiny number of binary features are required by our methods.

This paper is organized as follows. In Section 2 we introduce the linguistic phenomena used by annotators in the inference process, do a series of analyses on the human annotated dataset released by Mark Sammons et al., and point out five significant negative entailment phenomena. Section 3 specifies the five negative entailment phenomena in detail, proposes a rule-based method and a machine learning 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-1 task (Shima *et al*., 2011), annotate the dataset similar to the schema of Sammons *et al*. (2010), discuss if the negative entailment phenomena also appear in Chinese T-H pairs, and show their effects on TE in Chinese. Section 5 concludes the remarks.

2. Analyses of Human Inference Process in Textual Entailment

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 were chosen from the RTE-5 test set by Sammons *et al*. (2010), and total 39 linguistic phenomena divided into the following 5 aspects as follows, including knowledge domains, hypothesis structures, inference phenomena, negative entailment phenomena, and

[1] https://sites.google.com/site/ntcir11riteval/home-ct/task-guideline

knowledge resources, are annotated on the selected dataset. Table 1 summarizes the phenomena in the five aspects.

(a) **Knowledge Domains (Hypothesis Types)**: Each phenomenon in this aspect denotes whether the information in H belongs to the corresponding knowledge domain.

(b) **Hypothesis Structures**: Each phenomenon in this aspect denotes whether the H contains elements of the corresponding type.

(c) **Inference Phenomena**: Each phenomenon in this aspect indicates the corresponding linguistic phenomenon which is used to infer H from T.

(d) **Negative Entailment Phenomena**: Each phenomenon in this aspect is a pattern which may appear in negative entailment instances.

(e) **Knowledge Resources**: Each phenomenon in this aspect is a kind of knowledge or common senses which are required in the inference process in textual entailment.

Table 1. Five aspects of linguistic phenomena relating to textual entailment.

Aspect	Phenomena Types
Knowledge Domains	"be in", "cause", "come from", "create", "die/injure/kill", "group", "kinship", "name", "win/compete", "work"
Hypothesis Structures	"has Named Entity", "has Numerical Quantity", "has implicit relation", "has locative argument", "has nominalization relation", "has temporal argument"
Inference Phenomena	"coerced relation", "co-reference", "genitive relation", "implicit relation", "lexical relation", "nominalization", "passive-active", "wrong-label"
Negative Entailment Phenomena	"Named Entity mismatch", "Numeric Quantity mismatch", "disconnected argument", "disconnect relation", "exclusive argument", "exclusive relation", "missing modifier", "missing argument", "missing relation"
Knowledge Resources	"event chain", "factoid", "parent-sibling", "simple rewrite rule", "spatial reasoning", "numeric reasoning"

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. The implementation LIBSVM with the RBF kernel (Chang & 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 2 shows the results. Five aspects are first regarded as individual features, and then merged together. The two schemes, *Annotator* 1 and *Annotator* 2, mean the phenomena labelled by annotator 1 and annotator 2 are used as features, respectively. The scheme "1 *AND* 2", a strict criterion, denotes a phenomenon exists in a T-H pair only if both annotators agree with its appearance. In contrast, the scheme "1 *OR* 2", 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 features of 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 & Moruz, 2009). In negative entailment phenomena aspect, the "1 *OR* 2" scheme achieves the best accuracy whereas the performances of Annotator 1 and "1 OR 2" are the same in the setting with all the five aspects as features. In the following experiments, we adopt this labeling scheme.

Table 2. The accuracy of recognizing binary TE relation with the five aspects as features.

Aspect	Annotator 1	Annotator 2	1 AND 2	1 OR 2
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%

2.2 Negative Entailment Phenomena

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

There are nine linguistic phenomena in the aspect of negative entailment phenomena. We take each phenomenon as a single feature to do the task of two-way textual entailment recognition. Table 3 shows the experimental results. The first column is the phenomenon ID, the second column is the phenomenon, and the third column is the accuracy of using the

phenomenon in the binary classification. Comparing with the best accuracy 97.62% shown in Table 2, the highest accuracy in Table 3 is 69.52%, when missing argument is adopted. Each phenomenon may be suitable for some T-H pairs, and consequently all negative entailment phenomena together achieve the best performance.

Table 3. Accuracy of recognizing TE relation with individual negative entailment phenomena.

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%

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 the task of two-way entailment relation recognition by SVM classifiers. The notation C_n^m denotes a set of $m!/((m-n)!\times n!)$ feature settings, each with n features. For the sake of paper space, we only list the best 4 results in each combination set C_n^m shown in Table 4. Each feature setting is denoted by a set of phenomenon IDs enclosed parentheses. The notations between combination sets C_1^9~C_4^9 and C_5^9 ~C_8^9 are a slight difference because of the table space. For clarification, we list the phenomena not involved in the combination sets C_5^9 ~C_8^9. For example, the notation "-(0,1,2,6)" equals to the notation "(3,4,5,7,8)", which means the feature setting is composed of disconnected relation (ID: 3), exclusive argument (ID: 4), exclusive relation (ID: 5), missing argument (ID: 7) and missing relation (ID: 8).

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. On the other hand, removing some phenomena from (3, 4, 5, 7, 8) setting or adopting features rather than these phenomena decreases the performance. The best performance of using the feature

setting (-(0,6)), i.e., only 7 phenomena, is the same as that of using all 9 phenomena shown in Table 2.

Table 4. Accuracy of combination of negative entailment phenomena.

C_8^9		C_7^9		C_6^9		C_5^9	
-(6)	97.62%	-(0,6)	97.62%	-(0,1,6)	96.67%	-(0, 1,2,6)	95.24%
-(0)	97.62%	-(0,1)	97.14%	-(0,2,6)	96.19%	-(0,1,3,6)	94.29%
-(1)	97.14%	-(1,6)	96.67%	-(0,1,2)	96.19%	-(1,2,3,6)	93.33%
-(2)	96.67%	-(2,6)	96.67%	-(1,2,6)	95.71%	-(0,2,3,6)	93.33%
C_4^9		C_3^9		C_2^9		C_1^9	
(4,5,7,8)	92.38%	(4,7,8)	88.57%	(4,7)	79.52%	(7)	69.52%
(3,4,7,8)	91.43%	(3,4,7)	85.24%	(7,8)	79.05%	(8)	68.57%
(2,4,7,8)	90.48%	(0,7,8)	84.76%	(4,8)	78.57%	(4)	61.90%
(3,4,5,7)	90.00%	(4,5,7)	84.29%	(0,8)	76.67%	(0)	60.95%

We follow Sammons *et al.*'s definitions (2010) and describe the five significant negative entailment phenomena (3, 4, 5, 7, 8) as follows.

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

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

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

(d) **Missing Argument**: Entailment fails because an argument in H is not present in T, either explicitly or implicitly.

(e) **Missing Relation**: Entailment fails because a relation in H is not present in T, either explicitly or implicitly.

The correlations between these five phenomena are shown in Table 5. Each row presents the T-H pairs which are labelled with the corresponding negative entailment phenomenon by the scheme "1 OR 2". Each column in each row denotes the percentage of the T-H pairs which are also labelled with another negative entailment phenomenon. For example, the number of the T-H pairs which are labelled with "Disconnected Relation" is 14, and 2 of the 14 T-H pairs are also labelled with "Missing Argument". Therefore, the column "Missing Argument" in the

row "Disconnected Relation" shows the number 2/14 = 14.29%. Table 5 shows the low correlations between most significant negative entailment phenomena. In other words, these phenomena are complementary.

Table 5. Correlations between the five significant negative entailment phenomena.

	Disconnected Relation	Exclusive Argument	Exclusive Relation	Missing Argument	Missing Relation
Disconnected Relation	100.00%	0.00%	0.00%	14.29%	42.86%
Exclusive Argument	0.00%	100.00%	8.70%	8.70%	8.70%
Exclusive Relation	0.00%	16.67%	100.00%	0.00%	16.67%
Missing Argument	4.88%	4.88%	0.00%	100.00%	41.46%
Missing Relation	15.38%	5.13%	5.13%	43.59%	100.00%
Number of Occurrences	14	23	12	41	39

In the above experiments, we do all the analyses on the corpus annotated with linguistic phenomena by human. In some sense, we aim at knowing the ultimate performance of TE recognition systems embodying human knowledge in the inference. Of course, 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 achieved 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. Our experiments show the combination of these five phenomena is even more powerful. Vanderwende *et al.* (2006) suggested some phenomena that are the clue to false entailments. 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 is not always literal identical.

In the following, two methods are proposed to extract the phenomena from T-H pairs automatically in Section 3.2 and Section 3.3. The pre-processing of the pairs is described in Section 3.1.

3.1 Preprocessing

Before extraction, the English T-H pairs are pre-processed according to following considerations.

(a) **Numerical Character Transformation**: All the numerical values are normalized to a single format. The fractional numbers and percentages are converted to real numbers.

(b) **Stemming**: The stemming is performed to each word in the T-H pair with NLTK (Bird, 2002).

(c) **Part-of-Speech Tagging**: Stanford Parser is performed to tagging each word in the T-H pair (Levy & Manning, 2003).

(d) **Dependency Parsing**: Stanford Parser also generates the dependency pairs from T and H (de Marneffe *et al*., 2006). The results of dependency parsing contain crucial information for capturing negative entailment phenomena.

3.2 A Rule-Based Method

Noun phrases are the fundamental elements for comparing the existences of entailment. Given a T-H pair, we first extract 4 sets of noun phrases based on their POS tags: {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: {relation in H} and {relation in T}, where each relation in the sets is in a form of *Predicate*(*Argument1*, *Argument2*). 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 & 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 in T. Furthermore, we introduce WordNet to align synonyms in H and 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.

3.3 A Machine Learning Method

We aim at finding meta-features to describe the characteristic of negative entailment phenomena, and use them for classification. We analyse the dependencies in a T-H pair with Stanford dependency parser (de Marneffe *et al*., 2006) and derive two dependency sets DT and DH for T and H, respectively, where a dependency $gr(g,d)$ is in terms of a binary grammatical relation gr between a governor g and a dependent d. We further define the following three multisets to capture the relationships between T and H:

(a) {H only}$=\{gr|gr(g,d) \in D_H\text{-}(D_T \cap D_H)\}$

(b) {Partially identical in governor}$=\{gr|gr(g,d_1) \in D_T, gr(g,d_2) \in D_H, d_1 \neq d_2\}$

(c) {Partially identical in dependent}$=\{gr|gr(g_1,d) \in D_T, gr(g_2,d) \in D_H, g_1 \neq g_2\}$

A T-H pair is represented as a feature vector (V(a), V(b), V(c)), where the dimensions of the three vectors V(a), V(b), and V(c) are the number of grammatical relations in the

dependency parser. The weights of each grammatical relation *gr* in V(a), V(b), and V(c) are the number of *gr* appearing in the multisets {H only}, {Identical in governor only} and {Identical in dependent only}, respectively. The SVM classifier with the RBF kernel is adopted to develop classifiers with the parameters (cost and gamma) tuned by grid search and evaluated with 10-fold cross validation.

3.4 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 in the work of Mark Sammons *et al* (2010). 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 6 shows the performances of the negative entailment phenomena detection by rule-based and machine-learning methods. The performances of rule-based model are especially poor. The major challenge is to identify the arguments in T-H pairs. (S3) shows an instance. The correct arguments of H in (S3) are "Fifth Amendment right" and "driving license", but the arguments captured by our method are "Fifth Amendment" and "license". The issue can be improved with a better dependency parser.

(S3) **T**: "There is a rational basis to distinguish between people driving cars and semi trucks," Jambois said. "All I would say is I think he has an uphill battle." The lawsuit says the truckers' **Fifth** and Fourteenth **amendment rights** are being violated because there is no way for them to apply for an occupational license. Mutschler said the state is taking away the truckers' right to drive a truck for a living. He said he will argue that while **driving** is a privilege, once a person has a **license** for work, it becomes a right.
H: **Fifth Amendment right** is about **driving license**.

Table 6. Performance of negative entailment phenomena detection. Reported in Precision (P), Recall (R), and F-Score (F).

Aspect	Rule-based			Learning-based		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
Disconnected Relation	9.52	28.57	14.28	15.91	100.00	27.45
Exclusive Argument	12.94	47.83	20.37	15.49	95.65	26.66
Exclusive Relation	5.71	33.33	9.75	10.43	100.00	18.89
Missing Argument	32.11	37.81	34.72	38.46	97.56	55.17
Missing Relation	23.08	61.54	33.57	32.23	100.00	48.75
Average	16.67	41.82	22.54	22.50	98.64	35.38

Although the rule-based method is poorly-performed, and the machined learning method is not so good at precision and F-Score, the resulting models for TE recognition achieve decent performances. These interesting results are depicted in Table 7. The "Human-annotated" column shows the performance achieved by using the phenomena annotated by human. Using "Human-annotated" phenomena can be seen as the upper-bound of the experiments. In data set (a), the performance of using all the 5 phenomena as features by the machine learning method (M2) is better than that of using the rule-based method (M1). However, the results are reverse in data set (b). This may be because data set (b) contains some cases that cannot be recognized by the model trained from the T-H pairs annotated by human. On the other hand, the rule-based method is implemented directly from the definitions, which is more robust.

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 entailment recognition even though the phenomena detector is extremely simple. If we can improve the performance of the automatic phenomena detection algorithm, it may make a great progress on the textual entailment.

So far the experiments are two-stage classification. In the first stage, we perform the rule-based or the learning-based model to extract the five negative entailment phenomena. And then, the presences of the five phenomena are used as binary features to recognize the TE in the second stage. In this perspective, the features used for phenomena extraction in Section 3.3 are the *meta-features* of M2. In order to understand the impact of error-propagation, we train a one-stage TE recognizer, M3, by using the meta-features of M2 as features directly. Table 8 compares M1, M2, and M3. The models M2 and M3 do the TE recognition according to the same information, but the two-stage classifier M2 slightly outperforms M3. This result

suggests that the concept of negative entailment phenomena is useful for TE recognition.

Table 7. Accuracy of textual entailment recognition using the extracted phenomena as features.

	Dataset (a): 210 pairs			Dataset (b): 600 pairs	
	Rule-based (M1)	Learning-based (M2)	Human-annotated	Rule-based (M1)	Learning-based (M2)
Disconnected Relation	50.95%	54.76%	57.62%	54.17%	51.17%
Exclusive Argument	50.95%	50.95%	61.90%	55.67%	51.83%
Exclusive Relation	50.95%	52.38%	56.67%	51.33%	50.67%
Missing Argument	53.81%	57.62%	69.52%	56.17%	57.33%
Missing Relation	50.95%	50.95%	68.57%	52.83%	55.17%
All	52.38%	60.00%	95.24%	59.17%	57.83%

Table 8. Accuracies of two-stage and one-stage classification.

Stages of Classification	Model	Feature Source	Dataset (a): 210 pairs	Dataset (b): 600 pairs
Two-Stage	M1	Rule-based	52.38%	59.17%
	M2	Machine learning	60.00%	57.83%
One-Stage	M3	Meta-features of M2	56.19%	57.00%

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 dataset of RITE-1 BC-CT task in NTCIR-9. This dataset contains total 900 traditional Chinese T-H pairs, including 50% ENTAILMENT and 50% NO ENTAILMENT. We annotate all the nine 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. To avoid the influence from the actual entailment label (ENTAILMENT/NO ENTAILMENT), annotators can only see the part of T and H.

Table 9 shows the performances of TE recognition in Chinese with the human knowledge. The interpretation of this table is the same as that of Table 4. The accuracy of using all the nine 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 the same as those on English data. Besides, we can use only six phenomena to achieve the same performance as using all nine phenomena as features. Furthermore, we also classify the entailment relation by the phenomena extracted automatically by the rule-based method. The process is similar to those of English text described in Section 3.1 and Section 3.2, while Additional effort of processing is required for Chinese text. We segmented Chinese words with Stanford word segmenter (Chang *et al*., 2008) and performed Chinese dependency parsing using Stanford parser and the CNP parser (Chen *et al*., 2009). We extract two sets of negative entailment phenomena according to the parsing results of Stanford parser and CNP parser separately. Both sets are used as independent features to achieve a better performance.

Table 9. Accuracy of combination of negative entailment phenomena on Chinese data.

C_8^9		C_7^9		C_6^9		C_5^9	
-(1)	91.11%	-(1,6)	91.11%	-(1,2,6)	91.11%	-(0,1,2,6)	90.78%
-(2)	91.11%	-(1,2)	91.11%	-(0,1,2)	90.78%	-(1,2,3,6)	89.67%
-(6)	91.11%	-(2,6)	91.11%	-(0,1,6)	90.78%	-(1,2,6,8)	89.33%
-(0)	90.78%	-(0,1)	90.78%	-(0,2,6)	90.78%	-(0,2,4,6)	89.22%
C_4^9		C_3^9		C_2^9		C_1^9	
(3,4,5,7)	89.00%	(3,5,7)	86.11%	(3,7)	80.67%	(7)	74.89%
(3,5,7,8)	87.89%	(4,5,7)	84.78%	(5,7)	80.22%	(8)	67.89%
(0,4,5,7)	87.89%	(0,5,7)	84.67%	(4,7)	79.44%	(0)	56.89%
(1,3,5,7)	87.44%	(2,5,7)	83.89%	(0,7)	79.33%	(4)	56.67%

The rule-based method obtains a similar result of TE recognition in Chinese. 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 other methods using a lot of features, only 12 binary features are used in our method.

5. Conclusion

In this paper we conclude that the negative entailment phenomena have a great effect in dealing with TE recognition. The 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. To automatically capture the negative entailment phenomena in the text, we propose the phenomenon extraction algorithms with the rule-based and the learning-based approaches. 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 argument and relations, and determining their alignments need to be further examined to extract correct negative entailment phenomena. Besides, multi-class TE recognition will be explored in the future.

Reference

Androutsopoulos, I. & Malakasiotis, P. (2010). A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research, 38*, 135-187.

Bentivogli, L., Clark, P., Dagan, I., Dang, H. T., & Giampiccolo, D. (2011). The seventh PASCAL recognizing textual entailment challenge. In *Proceedings of the 2011 Text Analysis Conference* (*TAC 2011*), Gaithersburg, Maryland, USA.

Bentivogli, L., Dagan, I., Dang, H. T., Giampiccolo, D., & Magnini, B. (2009). The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the 2009 Text Analysis Conference* (*TAC 2009*), Gaithersburg, Maryland, USA.

Bird, S. (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)*, 69-72.

Chang, P.-C., Galley, M., & Manning, C. (2008). Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proceedings of the Third Workshop on Statistical Machine Translation* (*StatMT '08*), 224-232.

Chang, C.-C. & Lin, C.-J. (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.

Chen, W., Kazama, J., Uchimoto, K., & Torisawa, K. (2009). Improving Dependency Parsing with Subtrees from Auto-Parsed Data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (*EMNLP2009*), 570-579, Singapore.

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. *Lecture Notes in Computer Science*, 3944:177-190.

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (*LREC 2006*), 449-454.

Iftene, A. & Moruz, M. A. (2009). UAIC Participation at RTE5. In *Proceedings of the 2009 Text Analysis Conference* (*TAC 2009*), Gaithersburg, Maryland, USA.

Levy, R. & Manning, C. D. (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*), 439-446.

Sammons, M., Vydiswaran, V.G.V., & Roth, D. (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*), Uppsala, Sweden, 1199-1208.

Shima, H., Kanayama, H., Lee, C.-W., Lin, C.-J., Mitamura, T., Miyao, Y. *et al.* (2011). Overview of NTCIR-9 RITE: Recognizing inference in text. In *Proceedings of the NTCIR-9 Workshop Meeting*, Tokyo, Japan.

Vanderwende, L., Menezes, A., & Snow, R. (2006). Microsoft Research at RTE-2: Syntactic Contributions in the Entailment Task: an implementation. In *Proceedings of the Second PASCAL Challenges Workshop.*

Wang, R. & Zhang, Y. (2009). Recognizing Textual Relatedness with Predicate-Argument Structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 784–792.

Watanabe, Y., Miyao, Y., Mizuno, J., Shibata, T., Kanayama, H., Lee, C.-W. *et al*. (2013). Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. In *Proceedings of the 10th NTCIR Conference*, 385-404, Tokyo, Japan