NTU Textual Entailment System for NTCIR 9 RITE Task

Hen-Hsen Huang, Kai-Chun Chang, James M.C. Haver II, and Hsin-Hsi Chen
Department of Computer Science and Information Engineering,
National Taiwan University
Taipei, Taiwan
{hhhuang, kcchang, mchaver}@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

ABSTRACT

In this paper, we propose a system to deal with the Chinese textual entailment problem for NTCIR-9 RITE task. The RITE task consists of four subtasks, simplified Chinese binary classification (CS_BC), simplified Chinese multi-way classification (CS_MC), traditional Chinese binary classification (CT_BC), and traditional Chinese multi-way classification (CT_MC). According to the definitions of these subtasks, a machine learning based classification framework is proposed and tested under various setups. The performance of our system in the formal run achieves accuracies of 73.5%, 57.5%, 60.8%, and 48.3% for CS_BC, CS_MC, CT_BC, and CT_MC respectively.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Language Parsing and Understanding

General Terms

Algorithms, Experimentation, Human Factors, Languages

Keywords

Textual Entailment, Chinese Language Processing

1. INTRODUCTION

Recognizing Inference in Text (RITE), which is also known as Textual Entailment (TE), is an important task in Natural Language Processing. The task has been investigated in English for several years in PASCAL RTE challenges. Comparatively, there are still few researches in Chinese TE (neither traditional Chinese nor simplified Chinese) because of the lack of datasets and benchmarks. The RITE task in NTCIR-9 workshop provides a benchmark for researchers to evaluate the methods on this topic [1].

First we briefly explain what RITE is. Given a text pair, *Text* and *Hypothesis* denoted by t_1 and t_2, if one can consider t_2 (Hypothesis) is right by using the information of t_1 (or Text), we can infer t_2 from t_1 and say t_1 entails t_2 ($t_1 \rightarrow t_2$). Given two pairs of samples as follows:

(S1) “美國線上購併網景通訊公司。” ‘America Online acquired Netscape Communications.’

(S2) “網景被美國線上收購。” ‘Netscape is acquired by America Online.’

(S3) “川端康成是《雪國》的作者。” ‘Yasunari Kawabata is the writer of "Snow Country".’

(S4) “川端康成因他的小說《雪國》贏得了諾貝爾文學獎。” ‘Yasunari Kawabata won the Nobel Prize in Literature for his novel "Snow Country".’

The text in S1 can be inferred from the information of S2, and the text in S2 can be also inferred from S1. The text in S3 can be inferred from the information of S4, but S4 cannot be inferred from S3 because the information about “贏得了諾貝爾文學獎” (‘won the Nobel Prize in Literature’) is unavailable in S3.

In RITE task, there are three subtasks including binary-class (BC), multi-class (MC), and RITE4QA in three languages, Japanese, traditional Chinese, and simplified Chinese. We participate in the former two subtasks, i.e., MC and BC, in traditional Chinese and simplified Chinese. In binary-class subtask, each pair is labeled with (Y/N), where Y means t_1 entails t_2 and N means t_1 does not entails t_2. In multi-class subtask, each pair is labeled with one of relations including Bidirection, Forward, Reverse, Contradiction, and Independent. Forward (F) means t_1 entails t_2 ($t_1 \rightarrow t_2$), Reverse (R) means t_2 entails t_1 ($t_2 \rightarrow t_1$), Bidirection (B) means t_1 entails t_2 and vice versa, Contradiction (C) means we would consider t_2 is wrong when assume t_1 is right, and Independent (I) means we cannot tell whether t_1 entails t_2 nor do they exclude each other.

For example, the pair of (S1) and (S2) is a case of Bidirection because (S1) and (S2) entail each other. The pair of (S3) and (S4) is a case of Reverse because (S4) entails (S3) in one direction. The pair of (S5) and (S6) is Independent because the location information in (S5) “地點在中國” (‘sited in China’) is excluded from (S6), and the period information in (S6) “為期 184 天” (‘held for 184 days’) is excluded from (S5) as well. The pair of (S7) and (S8) is Contradiction because the state “宿敵” (‘competitor’) and the state “好朋友” (‘friend’) cannot be simultaneously true.

(S5) “昆明世界園藝博覽會地點在中國。” ‘Kunming International Horticultural Exposition is sited in China.’

(S6) “昆明世界園藝博覽會為期 184 天。” ‘Kunming International Horticultural Exposition is held for 184 days.’

(S7) “巴基斯坦的宿敵印度。” ‘Pakistan's competitor India.’

(S8) “巴基斯坦的好朋友印度。” ‘Pakistan's friend India.’

The released development dataset in traditional Chinese is 421 pairs in MC and 407 pairs in simplified Chinese. Because there are duplicate pairs in the two language datasets, we mainly used the traditional Chinese dataset to develop our system. The system is composed of two parts. We first extract features from t_1 and t_2 of all the pairs in the development dataset to train a SVM classifier after some lexical transformations like numerical normalization, and then introduce a rule-based post-classifier to improve the performance of the classifier.

In the development, we achieve an accuracy of 55.34% in traditional Chinese multi-class subtask and an accuracy of 74.35% in traditional Chinese binary-class subtask. In the formal run, our system has an accuracy of 48.3% in traditional Chinese multi-class subtask, 60.8% in traditional Chinese binary-class subtask, 57.5% in simplified Chinese multi-class subtask, and 73.5% in simplified Chinese binary-class subtask.

The rest of this paper is organized as follows. First, we review the related work in Section 2. In Section 3, our model along with the feature sets is introduced. The dataset used for the evaluation is illustrated in Section 4. In Section 5, the experimental results are discussed. Finally, we conclude this paper in Section 6.

2. RELATED WORK

There is little prior work related to Chinese textual entailment. In the following, we considered papers which deal with previous RTE challenges and English.

Androutsopoulos and Malakasiotis [2] review a variety of prior techniques for paraphrasing and recognizing textual entailment used on the MSR (Microsoft Research Paraphrase Corpus). The best results were achieved with surface string similarity. Besides, they describe all other techniques including logic-based approach, vector space models (VSM), syntactic similarity, symbolic meaning representation similarity, machine learning, and decoding. This provides a starting point when considering different techniques to apply as we can compare their efficacy.

Clark and Harrison [3] created a system composed of WordNet, DIRT and inference using a bag-of-words similarity model. Their results in RTE were 61.5% accuracy for two-way and 54.7% for three-way classification. By analyzing each step separately, they found the biggest gain was with WordNet and the least with parsing and DIRT.

Wang et al. [4] use a syntactic tree method comparison method to find similar questions within Yahoo! Answers. This particular method performs well within this area of research, so we decided to borrow the idea and apply it to textual entailment. Prior methods simply compared structural and syntactic similarities, but one also has to consider the role of semantics and production rules, so if two trees are similar but have different semantics or production rules there should still be some similarity.

Zhang and Zhang [5] use rule-based logic form transformation to solve question answering in Chinese and make experiments on Chinese TREC data. The main goals in the transformation are to find predicates, argument assignments, complex nominals, complex verbals, temporal representation, location representation, and propositional phrases, some of which are unique to Chinese.

Snow et al. [6] try to recognize false entailments through a combination of logical forms and syntactic heuristics. First they produce logical forms with NLPWin, which gives the syntactic dependencies. Then, they align the content nodes between the text and the hypothesis. The alignment is analyzed with the syntactic heuristics for false entailment. If none of them are true, then they compare them with a lexical similarity model. On RTE-1 they achieved 62.5% accuracy with this system.

3. THE MODEL

Our textual entailment system is based on the support vector machine (SVM) model. For each pair of texts, t_1 and t_2, we first construct the feature vector by finding the differences between t_1 and t_2 on several linguistic levels. The linguistic levels we considered include the word, the parsing tree, the dependency, the sentiment polarity, the referred name entity, and so on. In the training stage, all the pairs in the training set are transformed to the feature vectors and then sent to the SVM classifier. In the testing stage, each testing sample is also transformed to the feature vector in the same format and sent to the trained SVM model to determine the class of this sample. The resulting classes of a testing sample include Yes (Y, that is t_1 entails t_2) and No (N) in the binary classification and Bidirection (B), Forward (F), Reverse (R), Independent (I), and Contraction (C) in the multi-way classification.

In addition to the SVM classifier, a rule-based post-classifier is integrated into our system. The post-classifier performs a set of rules sequentially to make a final decision about the given sample. The rules are human-encoded with world knowledge that are hard to be captured by the machine.

3.1 Features

The features on various linguistic levels are utilized in our system.

Word (*W*): Both the texts in t_1 and t_2 are segmented into Chinese words by performing the Stanford Chinese Word Segmenter[i]. And then, three values are generated for each pair of t_1 and t_2. The first value is the number of the words appearing in t_1 only, the second value is the number of the words appearing in t_2 only, and the last value is the number of the words appearing in both t_1 and t_2.

Syntactic (*S*): Both the texts in t_1 and t_2 are also parsed to corresponding syntactic trees with the Stanford Parser[ii]. The fundamental structures of the syntactic tree are extracted as the features. The structure information includes the top three levels of the syntactic tree, the leftmost path of the syntactic tree, and the rightmost path of the syntactic tree. In addition, the subject, the major verb, and the object in a sentence are also considered as features.

Dependency (*D*): The dependency information between the elements on a sentence is also resulted from the Stanford Parser. All the pairs of the depended elements are treated as the elements in the bags. In this way, three bags of depended elements are constructed. The first bag includes the depended elements appearing on in t_1 only, the second bag includes the depended elements appearing on t_2 only, and the last bag includes the depended elements appearing on both t_1 and t_2. In addition, the numbers of each bag are also included.

Sentiment Polarity (*SP*): The sentiment polarity of each text (t_1 or t_2) is computed and included in the feature vector. The sentiment polarity detection is a hot issue in natural language processing itself. In this work, we implement a simple model that measures the sentiment polarity for a given text. In this model, the sentiment score of a text is computed by summarizing the individual sentiment score of each word in the text. The sentiment score is a positive value if the polarity of the text is positive, the score is a negative value if the polarity is negative, and the value denotes how strength the sentiment is. In this way, the sentiment polarity features include the polarity and the sentiment score of t1, the polarity and the sentiment score of t2, and whether the polarities of t1 and t2 are the same.

4. THE DATASET

The development data of the traditional Chinese tasks provided by the organizer are used as our training data and test data throughout this work. In this development data, there are 421 pairs of t_1 and t_2 that are labeled in 5-way for the multi-class subtask. We automatically build a 2-way dataset for the binary-class subtask by labeling all the Forward/Bidirection samples in 5-way as Yes and labeling all the Reverse/Contradiction/Independent samples in 5-way as No.

After the formal run, the organizer published four sets of the testing data used in the formal run. The four sets include CS_BC, CS_MC, CT_BC, and CT_MC corresponding to four Chinese subtasks, simplified Chinese in binary-class, simplified Chinese in multi-class, traditional Chinese in binary-class, and traditional Chinese in multi-class. We also test our model with these four datasets.

The statistics of experimental datasets are shown in Table 1 and Table 2. The labeling in the development dataset is sometimes confusing. For instance, the pair of (S9) and (S10) from the sample 10035 is labeled as Reverse by the annotator. However, the information about “帶來台灣” (‘brought Zhang Xue-Liang to Taiwan’) in (S9) is completely unavailable in (S10). Thus, the pair of (S9) and (S10) is actually a case of Independent.

(S9) “蔣介石將張學良帶來台灣軟禁。” ‘Chiang Kai-Shek brought Zhang Xue-Liang to Taiwan and house-arrested him’

(S10) “張學良、孫立人、彭明敏等人，都是遭蔣介石軟禁的著名例子。” ‘Zhang Xue-Liang, Sun Li-Jen, and Peng Ming-Min are the notorious examples that they were under house arrest by Chiang Kai-shek.’

We still keep the original labels because the testing data used in the formal run are also annotated by the organizer.

Table 1. Statistics of the dataset for binary-class subtask

Dataset	Y	N	Total
Development	169	252	421
CS_BC	263	144	407
CT_MC	450	450	900

Table 2. Statistics of the dataset for multi-class subtask

Dataset	B	F	R	C	I	Total
Development	82	87	97	74	81	421
CS_MC	71	101	91	74	70	407
CT_MC	180	180	180	180	180	900

5. EXPERIMENTS

The experiments are set up as two parts. In the first part, we evaluate our model with different feature sets used in the development dataset. In this stage, the performance of each feature set and the performance of the combination of feature sets are revealed. Ten-fold cross-validation is applied in this stage. In the second part, the best model trained from the first part is chosen to be tested under the four testing datasets from the formal run.

5.1 Evaluation Metrics

All the evaluation performances are reported in terms of accuracy, precision, recall, and F-score.

Accuracy measures how many samples are correctly predicted. For each class, we define precision as the ratio of the correctly predicted samples to all the samples predicted as the class, recall as the ratio of the correctly predicted samples to all the samples that actually belong to the class, and F-score as the harmonic mean of precision and recall.

5.2 Feature Performance

The performance of each feature set in binary classification is shown in Table 3, and the performance of each feature set in multi classification is shown in Table 4. In both subtasks, the word level features are the best performed feature set. With the combination of all the features, the performances slightly improve to 74.58% and 55.58% in the subtasks binary-class and multi-class respectively.

The confusion matrices of the best models for both subtasks are given in Table 5 and Table 6. Each row in the matrices is the distribution of the classification result for a class. For instance, there are 148 samples that are correctly classified as Yes and 21 samples that are wrongly classified as No in Table 5. In binary-class subtask, the model tends to estimate input samples as Yes. In multi-class subtask, the performance is relatively poor for the classes Contradiction and Independent. The cases of Contradiction tend to be misclassified as Bidirection, Forward, and Reverse. The cases of Independent tend to be misclassified as Forward and Reverse.

Table 3. Feature Performance in Binary-Class

Features	Accuracy	Precision	Recall	F-Score
Word	74.11%	76.71%	76.91%	74.11%
Syntactic	57.72%	56.62%	56.79%	56.62%
Dependency	71.50%	72.41%	73.17%	71.40%
Sentiment	39.90%	36.64%	49.61%	28.84%
All	74.58%	76.01%	76.72%	74.54%

Table 4. Feature Performance in Multi-Class

Features	Accuracy	Precision	Recall	F-Score
Word	52.02%	44.64%	49.50%	44.15%
Syntactic	27.79%	29.62%	27.69%	27.73%
Dependency	51.07%	40.08%	48.16%	39.32%
Sentiment	19.24%	3.89%	19.76%	6.51%
All	55.58%	53.37%	53.74%	51.56%

Table 5. Confusion Matrix of Binary-Classification

Class	Y	N	Precision	Recall	F-Score
Y	148	21	63.25%	87.57%	73.45%
N	86	166	88.77%	65.87%	75.63%
Overall	234	187	76.01%	76.72%	74.54%

Table 6. Confusion Matrix of Multi-Classification

Class	B	F	R	C	I	Precision	Recall	F-Score
B	48	9	11	11	3	51.61%	58.54%	54.86%
F	7	70	1	3	6	58.33%	80.46%	67.63%
R	9	2	75	2	9	60.48%	77.32%	67.87%
C	20	17	15	15	7	45.45%	20.27%	28.04%
I	9	22	22	2	26	50.98%	32.10%	39.39%
Overall	93	120	124	33	51	53.37%	53.738%	51.558%

5.3 Formal Run Performance

We train our models with all features and all development data for the formal run. For each subtasks, three results are submitted. The first result is estimated with the SVM classifier and the post-classifier. The second result is estimated with the SVM classifier and the post-classifier under a different parameter setting. The third result is estimated with the SVM classifier only. The best results of the formal run are given in Table 7. We train the model based on the development data for traditional Chinese subtask. However, our model achieves better performances in simplified Chinese subtasks.

Table 7. Formal Run Performance

Subtask	Accuracy	Precision	Recall	F-Score
CS_BC	73.46%	70.96%	70.83%	70.89%
CS_MC	57.49%	56.07%	54.42%	53.08%
CT_BC	60.78%	62.86%	60.78%	59.12%
CT_MC	48.33%	47.08%	48.33%	46.07%

6. CONCLUSION

In NTCIR-9 RITE task, we explore Chinese textual entailment in both binary-class and multi-class subtasks. We proposed a hybrid approach that combines a learning-based SVM classifier and a rule-based post classifier to deal with this problem. In the development stage, our model achieves an accuracy of 55.34% in traditional Chinese multi-class subtask and an accuracy of 74.35% in traditional Chinese binary-class subtask. In the formal run, our model achieves accuracies of 73.5%, 57.5%, 60.8%, and 48.3% in the subtasks simplified Chinese binary-class, simplified Chinese multi-class, traditional Chinese binary-class, and traditional Chinese multi-class respectively.

7. REFERENCES

[1] H. Shima, H. Kanayama, C.-W. Lee, C.-J. Lin, T. Mitamura, Y. Miyao, S. Shi, and K. Takeda. 2011. Overview of NTCIR-9 RITE: Recognizing Inference in TExt. In *NTCIR-9 Proceedings*, to appear.

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

[3] Clark, Peter and Harrison, Phil. 2009. An Inference-based Approach to Recognizing Entailment. In *Proceedings of TAC.*

[4] Wang, Kai, Ming, Zhaoyan, Chua, Tat-Seng. 2009. A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services. In *Proceedings of SIGIR*.

[5] Zhang, Yi and Zhang, Dongmo. 2003. Enabling Answer Validation by Logic Form Reasoning in Chinese Question Answering. In *Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2003)*, Beijing, China.

[6] Snow, Rion, Vanderwende, Lucy, and Menezes, Arul. 2006. Effectively Using Syntax for Recognizing False Entailment. In *Proceedings of the North American Association of Computational Linguistics*, New York City, New York, United States of America.

[i] http://nlp.stanford.edu/software/segmenter.shtml

[ii] http://nlp.stanford.edu/software/lex-parser.shtml