Extracting Features for Machine Learning in NTCIR-10 RITE Task

Lun-Wei Ku
Institute of Information Science, Academia Sinica
128 Academia Road, Section 2
Nankang, Taipei 115, Taiwan
+886-2-27883799 ext 1808
lwku@iis.sinica.edu.tw

Edward T.-H. Chu
National Yunlin University of Science and Technology
123 University Road, Section 3,
Yunlin 64002, Taiwan
+886-5-5342601
edwardchu@yuntech.edu.tw

Nai-Hsuan Han
National Yunlin University of Science and Technology
123 University Road, Section 3,
Yunlin 64002, Taiwan
+886-5-5342601
m10017007@yuntech.edu.tw

ABSTRACT
NTCIR-9 RITE task evaluates systems which automatically detect entailment, paraphrase, and contradiction in texts. We developed a preliminary system for the NTCIR-9 RITE task based on rules. In NTCIR-10, we tried machine learning approaches. We transformed the existing rules into features and then added additional syntactic and semantic features for SVM. The straightforward assumption was still kept in NTCIR-10: the relation between two sentences was determined by the different parts between them instead of the identical parts. Therefore, features in NTCIR-9 including sentence lengths, the content of matched keywords, quantities of matched keywords, and their parts of speech together with new features such as parsing tree information, dependency relations, negation words and synonyms were considered. We found that some features were useful for the BC subtask while some help more in the MC subtask.

Categories and Subject Descriptors
I.2.7 [**Natural Language Processing**]: *Language parsing and understanding, Text analysis*

General Terms
Algorithms.

Keywords
NTCIR-10 RITE, POS features, semantic features, syntactic features, SVM.

Team Name
Yuntech

Subtasks
BC (CT, CS), MC (CT, CS)

1. Introduction
RITE [1] is a generic benchmark task that addresses major text understanding needs in various NLP/Information Access research areas. It evaluates systems which automatically detect entailment, paraphrase, and contradiction in texts written in Japanese, simplified Chinese, or traditional Chinese. Entailment is a classic logic problem in the research domain of artificial intelligence, and in the RITE task it becomes a natural language processing problem while the experimental materials are texts.

We believe that to achieve more fruitful results, incorporating the learning methods by a hybrid approach or using linguistic clues in a machine learning approach might be necessary. In NTCIR-9, our first attempt was adopting rules. In NTCIR-10, we tried to extend these rules and to transform them into several features. Sentence lengths, part of speech tags, and matched keywords were the major features in the designed rules in NTCIR-9. New features in NTCIR-10 including parsing information, dependency relations, negations and synonyms were also utilized. Furthermore, as we want to identify the true different parts between two sentences, the preprocessing was enhanced by using consistent numeric expressions and units. The developed system was for the binary classification (BC) and multiple classification (MC) subtasks on both traditional (TC) and simplified (SC) Chinese materials. We submitted two runs. In the following sections, we will first describe our system flow, discuss the extracted features, and then analyze the performance of each feature. We will also show the comparison of our performance in NTCIR-9 and NTCIR-10. At last, a best feature set is proposed.

2. System Description
We adopted support vector machine (SVM) as the machine learning algorithm. Figure 1 shows the system flow. The test sentences, T1 and T2, were first preprocessed. Sentences were segmented and parsed by the CKIP parser [2], and also parsed by the Stanford parser to get the dependency relations [5]. Numeric words and units of measurement were unified into the same form. After that, features were extracted from the context of two sentences, and additional resources such as Cilin which provided Chinese synonyms and a negation word list were incorporated to get more features. Four kinds of features were utilized as shown in Figure 1: POS feature denotes the part of speech feature; Rule feature denotes the feature inspired by the rules we adopted in NTCIR-9; Statistic feature denotes the feature related to a number of occurrence or an average value; Syntactic feature denotes the feature from parsed sentences. SVM then used these features to predict the relation between each testing sentence pair. The extracted features were for both BC and MC subtasks.

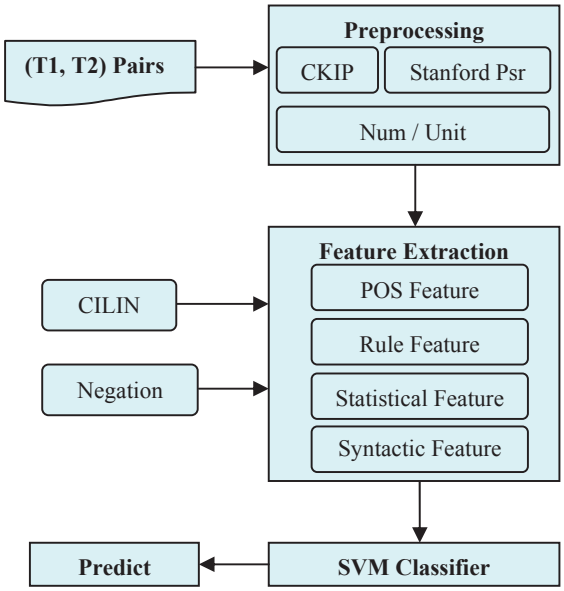

Figure 1. System flow

3. Extracting Features for SVM

As mentioned, preprocessing was performed before extracting features. We participated in the traditional and simplified Chinese tasks, for which segmentation was necessary to know the word boundaries. Part of speech tagging was done after the segmentation [3]. With parts of speech, we can make numeric words and units of measurement into the same form. All numeric expressions were transformed into Arabic numbers. For example, 2,000 萬 (ten thousands) was transformed into 20,000,000. Timestamps which were of part of speech Nd, were also changed into the same expression form. For example, 一九九七年 was changed into 1997 年 (year). In addition, units of measurement, which were of part of speech Nf, were also standardized. For example, 公尺 and 米 were both changed into m. After the preprocessing, identical words were removed before feature extraction. As the principle of feature extraction is to find out the identical and different parts between two sentences, these unifications can help to get more accurate features before machine learning.

In this section, we will describe features and the performance with and without them. The results shown in this section are from 5-fold experiments on the RITE2 training data.

3.1 Sentence Length

In NTCIR-9, there was one rule which divided sentence pairs into two categories: |T1-T2| < 2 and |T1-T2|>=2. That is, we thought that for sentence lengths, a difference less than two could not provide additional information and therefore made two sentences identical (bi-direction). In NTCIR-10, we generated features to represent three categories instead of two. We add an additional uncertain zone when the difference of length is between two to five. In other words, when the difference was lower than 2, we were pretty sure they were identical; higher than 5, not identical; but from 2 to 5, other features might help to determine their relation better. The threshold value 5 is from the average difference between independent sentences in NTCIR-9. The performance of with and without these features were shown in Table 1 (Length Features#1). Features for baseline were extracted after removing identical words between two sentences. Features extracted from sentence length introduced in this section and parts of speech in next section were included.

However, we found that a small difference of sentence length did not necessary mean that two sentences were similar. On the contrary, if many words were left after removing identical words, they did express different information in two sentences. Therefore, in addition to the difference of sentence length, we checked the number of left words in the shorter sentence. Moreover, the percentage of words left after removing identical words was also used as a feature. In total, we designed seven additional features of coverage conditions related to the sentence length as shown in Figure 2 (Length Features#2).

1. T1 > T2 and T2 = 0
2. T2 > T1 and T1 = 0
3. T1 = T2 = 0
4. T2 = T2 ≠0
5. T2 ≠ T2 ≠0
6. T1 (identical words removed) / T1 (original)
7. T2 (identical words removed) / T2 (original)

Figure 2. Coverage conditions

Table 1. Accuracy of baseline w/o Length Features#1

Subtask	Baseline (%)	w/o Length Features#1 (%)
BC (CS)	72.73	65.97
BC (CT)	63.06	63.43
MC (CS)	62.29	51.72
MC (CT)	54.58	48.22

Table 2. Accuracy of baseline w/o Length Features#2

Subtask	Baseline (%)	w/o Length Features#2 (%)
BC (CS)	72.73	71.50
BC (CT)	63.06	61.77
MC (CS)	62.29	60.07
MC (CT)	54.58	53.75

3.2 Parts of Speech

Parts of speech provided us important information of the roles words in one sentence played. In NTCIR-9, we considered whether verbs (Va, Vc, Vh, Vj), nouns (Nb, Nc, Nd), numbers (Neu), adverbs (D) and location conjunctions followed locations (VCL+Nc) were identical in two sentences. In NTCIR-10, we further checked whether their quantities and indexes were identical. Table 3 shows the performance of using these features and not using them. Furthermore, Figure 3 and Figure 4 show the effect of each part of speech feature in traditional BC and MC subtasks. These two figures show that all POS features benefit the MC subtask while part of speech VA, Nc and Ncd deteriorate the performance of the BC subtask.

Table 3. Accuracy of baseline w/o all POS features

Subtask	Baseline (%)	w/o POS Features (%)
BC (CS)	72.73	72.00
BC (CT)	63.06	64.35
MC (CS)	62.29	60.93
MC (CT)	54.58	55.79

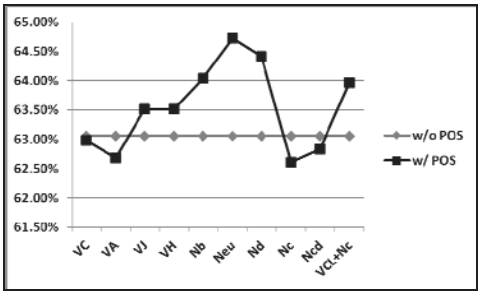

Figure 3. The effect of each POS feature (BC (CT))

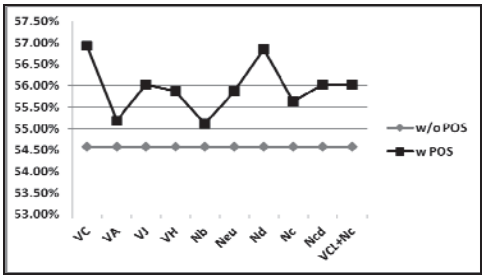

Figure 4. The effect of each POS feature (MC (CT))

3.3 Important Entities

From previous research, we have known that people, event, time, location, objects are important clues to understand the content of a sentence. In NTCIR-9, we considered them in a subtle way, which extracted words with corresponding parts of speech and generated rules. Among these clues, event is a more complicated concept and difficult to be extracted correctly compared to others, while objects were already included by the nouns we have considered in section 3.2. In NTCIR-10, features corresponding to people, time, and location were considered for machine learning. To extract them, words indicating people whose part of speech was Nb, words indicating time whose part of speech was Nd, and word pairs indicating locations whose parts of speech were Nc plus Ncd were left for comparisons. The same mechanism for extracting sentence length features was applied here again and five features of coverage conditions were as shown in Figure 2, the first five.

Table 4, 5 and 6 show the effect of People, Time and Location features, respectively. Performance shows that Time and Location features benefit more than People features. However, we don't know yet whether it is because more Time and Location related words are mentioned in the training set.

Table 4. Accuracy of baseline w/o People features

Subtask	Baseline (%)	w/o People Features (%)
BC (CS)	72.73	72.73
BC (CT)	63.06	63.74
MC (CS)	62.29	62.16
MC (CT)	54.58	55.19

Table 5. Accuracy of baseline w/o Time features

Subtask	Baseline (%)	w/o People Features (%)
BC (CS)	72.73	72.97
BC (CT)	63.06	63.44
MC (CS)	62.29	63.51
MC (CT)	54.58	54.73

Table 6. Accuracy of baseline w/o Location features

Subtask	Baseline (%)	w/o People Features (%)
BC (CS)	72.73	72.11
BC (CT)	63.06	63.59
MC (CS)	62.29	61.92
MC (CT)	54.58	56.17

3.4 Syntactic Features

To improve the performance, we tried to utilize syntactic features in NTCIR-10. We hope the features related to phrases and sub-sentences can help classify the relation between two sentences. In this paper, parsing information was the source of syntactic features, which were from the sentence parsing trees and the dependency relations. We considered syntactic structures related to important entities mentioned in the previous section. Therefore, we utilized parsing trees by comparing VP (verb phrase) and NP (noun phrase) subtrees of two sentences and found the identical ones. Among these identical subtrees, we located the one whose head was of the lowest depth and used the inverse of this depth (one mod depth) as the feature value. This value can represent the coverage of the biggest identical subtree. The bigger the coverage is, the bigger the subtree is so that the larger the inverse of the depth is. The largest value of this feature is 1. Table 7 shows the performance of adding the inverse of parsing tree depth as a feature.

Table 7. Accuracy of baseline w/ the Inverse of Parsing Tree Depth features

Subtask	Baseline (%)	w/o People Features (%)
BC (CS)	72.73	72.85
BC (CT)	63.06	63.51
MC (CS)	62.29	62.65
MC (CT)	54.58	54.43

Dependency relations represent syntactic structures of a parsing tree in a different way. We obtained the dependency relations by processing sentences with the Stanford parser [4]. Relations related to important entities mentioned in section 3.3, i.e., Obj, Subj, NUM, TMOD, DEP and NN, were analyzed. We compared the relations of two sentences and dropped the identical ones. Then we determined the coverage conditions of important entities again by features as shown in Figure 2, the first five ones. Table 8 shows the performance of adding dependency relations as features, and Figure 5 shows the effect of each dependency relation. From Figure 3 and Figure 5, we can see that not all dependency relations we selected were helpful, while those POS we selected were. Figure 3, 4, 5, and 6 also reveal that useful features for BC and MC subtasks could be different.

Table 8. Accuracy of baseline w/ Dependency Relations features

Subtask	Baseline (%)	w/o People Features (%)
BC (CS)	72.73	71.74
BC (CT)	63.06	64.72
MC (CS)	62.29	60.81
MC (CT)	54.58	54.73

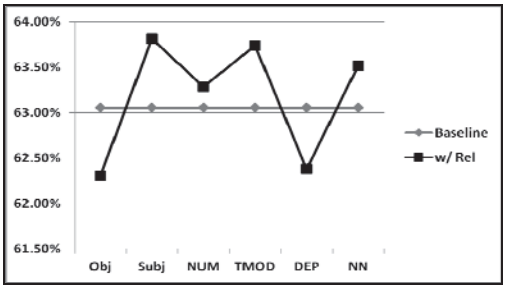

Figure 5. The effect of Dependency Relation features (BC (CT))

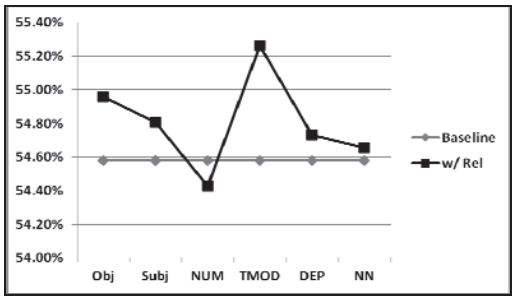

Figure 6. The effect of Dependency Relation features (MC (CT))

3.5 Negations and Synonyms

Negations and synonyms are the semantic features used in our system. Negations are features with which we tried to deal with simple conflict sentences. Three features were related to negations: negation appearance (true/false), number of negations, and whether two sentences contain the same negations (true/false). The effect of these three features is shown in Table 9. Surprisingly, the performance dropped after adding these features. It was because negations did appear in sentences of conflict relations, but not only in sentences of conflict relations. Therefore, the negation word itself cannot tell the relation of sentence pairs. Other features must be utilized together with it.

Table 9. Accuracy of baseline w/ Negation features

Subtask	Baseline (%)	w/o People Features (%)
BC (CS)	72.73	69.78
BC (CT)	63.06	61.01
MC (CS)	62.29	61.06
MC (CT)	54.58	55.34

The second semantic features, synonyms, were utilized to identify semantically identical words. Cilin [6] was used to expanded synonym words for matching. In addition to the appearance of synonyms (true/false), synonym related features were obtained after removing semantically identical words of two sentences. Five features were extracted as in Figure 2, the first five. Table 10 shows the effect of synonym features. We can see that the synonym features averagely benefit the performance.

Table 10. Accuracy of baseline w/ Synonym features

Subtask	Baseline (%)	w/o People Features (%)
BC (CS)	72.73	72.73
BC (CT)	63.06	63.97
MC (CS)	62.29	62.65
MC (CT)	54.58	55.41

4. RITE Task Performance and Discussion

Table 11 shows the overall performance of adopting features in Section 3, where I denotes the baseline, P the syntactic features, N the negation features, and C the synonym features. From the experiments on the training set, I+P+N+C was the best feature set. In NTCIR-10 RITE2, we submitted three runs, and the feature set of run 1 was I+P, run 2 I+P+N, and run 3 I+P+N+C. Table 12 shows the performance of three submitted runs, where run 3 was of the best performance, which conformed with the results of experimenting on the training set. Table 13 shows the performance of our rule-based system in NTCIR-9 for comparison. As the training set in NTCIR-10 is the testing set of NTCIR-9, from Table 11 and Table 13, we can see that when we transformed rules into features of machine learning, the average performance of four subtasks improved 0.09 (16.54%). When we further added additional syntactic and semantic features, the performance improved 0.10 (19.31%).

Table 11. Accuracy of experimenting on the training set of NTCIR-10

Subtask	I	I+P	I+P+N	I+P+N+C
BC (CS)	**72.73**	71.74	71.01	71.13
BC (CT)	63.06	63.82	64.35	**65.71**
MC (CS)	62.29	60.44	61.18	**62.53**
MC (CT)	54.58	57.31	57.38	**58.44**

Table 12. Accuracy of Run 1, Run 2 and Run 3 for the BC and MC subtasks in NTCIR-10

Run \ Task	BC (CS)	BC (CT)	MC (CS)	MC (CT)
1	58.70	62.47	50.77	51.81
2	59.08	62.59	51.28	51.36
3	59.59	62.59	51.28	52.04

Table 13. Accuracy of Run 1 and Run 2 for the BC and MC subtasks in NTCIR-9

Run \ Task	BC (CS)	BC (CT)	MC (CS)	MC (CT)
1	63.60	52.80	52.80	47.70
2	56.00	52.40	39.80	38.80

5. Error Analysis

To analyze our systems, Table 14 and 15 show the confusion matrices of the performance of BC and MC subtasks. Table 14 shows that the system did better in NTCIR-10 than in NTCIR-9 for negative sentence pairs.

Table 14. Confusion matrix (RITE2-Yuntech-CT-BC-run3)

	Y	N	
Y	310	169	479
N	161	241	402
	471	410	

Table 15. Confusion matrix (RITE2-Yuntech-CT-MC-run3)

	F	B	C	I	
F	297	17	3	11	328
B	37	78	12	24	151
C	67	28	3	16	114
I	181	22	5	80	288
	582	145	23	131	

Table 15 shows that we had the worst performance in the contradiction category. After taking a closer look at sentences in this category, we found that the resources we utilized were not enough for identifying these sentences. To identify them, we need to know antonyms. For example, in the following sentence pair,

S1: 草莓是薔薇科草莓屬植物中最**常見(common)**的一種

S2: 草莓是薔薇科草莓屬植物中最**稀有(rare)**的一種

, we need to know common and rare are antonyms to categorize them correctly. Moreover, some words are not antonyms but they express conflict concepts. For example, in the following sentence pair,

S1: 孔子早年的**生活(life) 極為(extremely) 艱辛(difficult)**

S2: 孔子的早期的**生活(life) 十分(very) 優渥(wealthy)**

, the words difficult and wealthy represent conflict concepts only when they are describing certain things like life. To know this, our system first needs to learn knowledge from a lot of data. Adding negations as features for the conflict category as we did is not enough. As a result, this sentence pair was reported independent instead of conflict.

In addition to the lacking of resources, we found that some features were very obvious but not functional because they were dominated by other features. For example, in the following sentence pair,

S1: **外國人(foreigners, word 1)** 尊稱(word 2) **孫中山(Sun yat sen, word 3)** 為現代中國之父

S2: **孫中山(Sun yat sen, word 1)** 尊稱(word 2) **外國人(word 3)** 為現代中國之父

, though we considered the position of words as a feature, more other features dominated it while these two sentences were very similar so that this pair was reported bidirectional. Moreover, in the following sentence pair,

S1: 美國地區於前年 11 月**傳出(reported)** 狂牛症病例衛生署去年 6 月再度禁止美國牛肉輸來台灣

S2: 美國**未(not) 傳出(reported) 任何(any)** 狂牛病例

, a portion of information in the first sentence conflicted with the second sentence. However, the information in the first sentence was obviously much more than in the second one. To deal with this kind of sentence pairs, the system might need the ability to match fragments of two sentences. This pair was misjudged forward also because other features dominated the key feature. This phenomenon suggested that, as we expected, having a hybrid system of rule based and machine learning may enhance the performance.

Another problem we might be able to fix was the unification of digits plus units of measurements. For example, zero Celsius degree equals thirty two Fahrenheit degree. We only transformed the units of measurements but not considering digits at the same time. Therefore, we failed to process sentences like the following:

哈利法塔高度為 **828 公尺(meters)**

哈利法塔高度為 **828 英尺(feets)**.

6. Conclusion and Future Work

We have designed a preliminary rule-based system in NTCIR-9 which considered the different parts of a sentence pair to tell whether the first one can infer the second one or to determine the type of sentence relations. In NTCIR-10, we transformed these rules to features and added additional syntactic and semantic features for a machine learning system. More complete preprocessing steps, such as numbers, time and units standardizations were also performed in this system.

According to the experiment results, we found that using all the proposed features, i.e., parts of speech, syntactic and semantic features, improves the performance the most. However, according to the analysis, we also found some features individually did not help in telling the sentence relations. Therefore, we might still have room to improve the system by advanced feature selection skills.

As we have tried a rule based system in NTCIR-9 and a machine learning system in NTCIR-10, we will develop a hybrid system in the future to achieve a better performance. In NTCIR-10, we only utilized parsing trees and synonyms as features. More deep analyses and comprehensive comparisons of sentence structures and word senses are also our future work.

7. ACKNOWLEDGMENTS

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC 101-2628-E-224-001-MY3.

8. REFERENCES

[1] H. Shima, H. Kanayama, C.-W. Lee, C.-J. Lin, T. Mitamura, Y. Miyao, S. Shi, and K. Takeda. Overview of NTCIR-9 RITE: Recognizing Inference in TExt. In NTCIR-10 Proceedings, to appear, 2013.

[2] CKIP Chinese word segmentation system. http://ckipsvr.iis.sinica.edu.tw/

[3] CKIP (Chinese Knowledge Information Processing Group). (1995/1998). *The Content and Illustration of Academica Sinica Corpus*. (Technical Report no 95-02/98-04). Taipei: Academia Sinica.

[4] Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? ACL 2003, pp. 439-446.

[5] Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative Reordering with Chinese Grammatical Relations Features. In Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation.

[6] Mei, J., Zhu, Y. Gao, Y. and Yin, H.. (1982). tong2yi4ci2ci2lin2. Shanghai Dictionary Press.