The Yuntech System in NTCIR-9 RITE Task

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

Lun-Wei Ku
National Yunlin University of Science and Technology
123 University Road, Section 3, Douliou,
Yunlin 64002, Taiwan
+886-5-5342601
lwku@yuntech.edu.tw

ABSTRACT
NTCIR-9 RITE task evaluates systems which automatically detect entailment, paraphrase, and contradiction in texts. The Yuntech team developed a preliminary system for the NTCIR-9 RITE task and was described in this paper. The major aim of this system was to determine the type of the relation of two sentences. A straightforward assumption was proposed for achieving this aim: the relation between two sentences was determined by the different parts between them instead of the identical parts. Therefore, we considered features including sentence lengths, the content of matched keywords, quantities of matched keywords, and their parts of speech to capture the difference between two sentences. Rule-based methods were implemented to develop the system according to the proposed assumption, and good performances were achieved for some types.

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

General Terms
Algorithms.

Keywords
NTCIR-9 RITE, entailment, paraphrase, contradiction, rule-based.

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.

To solve a natural language processing research problem, using linguistic clues and rules are often the first attempt. It can usually give a passable result for the research problem. However, to achieve more fruitful results, we may need to incorporate the learning methods by a hybrid approach or use linguistic clues in a machine learning approach.

In the beginning of exploring the research problem of determining whether the first sentence could infer the second (the BC subtask) or detecting the types of sentence relations (which refer to forward entailment, paraphrase, reverse entailment, contradiction, and independence here after; the MC subtask), linguistic clues and rules were adopted to construct the Yuntech preliminary system for the RITE task. As participating in this task in a very late time, a straightforward assumption was proposed to achieve this aim: the relation between two sentences was determined by the different parts between them instead of the identical parts. Under this assumption, sentence lengths, part of speech tags, and matched keywords were adopted as the major features in the designed rules. The developed system was used in the BC and MC subtasks, experimenting on both traditional (TC) and simplified (SC) Chinese materials and submitting two runs. In the following sections, we will first describe our system flow, show the experimental results, and then discuss the issues of the proposed method.

2. System Description
We designed a simple linguistic rule-based model for RITE as we hoped to develop a system as quick as possible to meet the schedule. Figure 1 and Figure 2 show the flows of the system for the BC and MC subtasks in Run 1. Sentences were segmented by the CKIP segmentation system before processing [2]. We defined some criteria for judging the types of sentence relations as shown in system flows. For the BC subtask, the criteria include:

1. Whether any word of the part of speech (POS) VA, VC, VH, VJ (verbs), and Nb (proper names) [3] appearing in T1 is the same to word of the corresponding POS in T2.
2. Whether the word pair of POS VCL (verb followed by location object) and Nc (location noun) in T1 is the same to that in T2.

For the MC subtask, the criteria include:

1. Whether any word of the part of speech (POS) Nb, Nc, Nd, Neu (numbers), Ncd, VA, VC, VH, VJ (verbs) [3] in T1 is identical to the word of the corresponding POS in T2.
2. Whether the number of negation words are odd or even in both T1 and T2.

The criterion I in Figure 2 is the rule to identify the independence type (I). If T1 and T2 contain different proper names, and they both include or do not include adverbs (D), they are determined as independent by the system.

The flows of the system in Run 2 are slightly different from those in Run 1. Because the answers from the systems for the BC and MC subtasks in Run 1 were generated independently, for the BC subtask in Run 2 we utilized the MC results of Run 1 to give answers directly. The mapping rules will be introduced in section

3. For the MC subtask in Run 2, an additional criterion was added in the gray rhombus:

3. Whether the number of identical words of the POS selected in criterion 1 exceeded 5.

Adding this criterion in MC will add weights to the identical parts of two sentences.

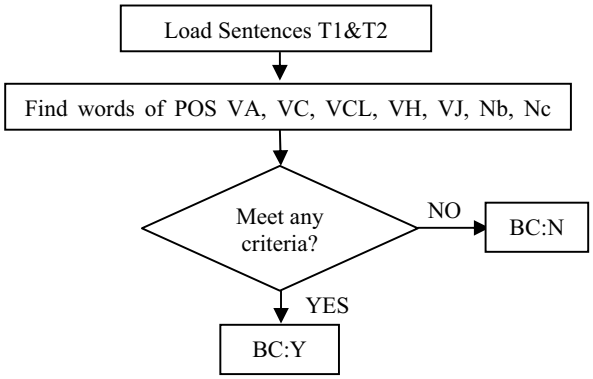

Figure 1. System flow for the BC subtask

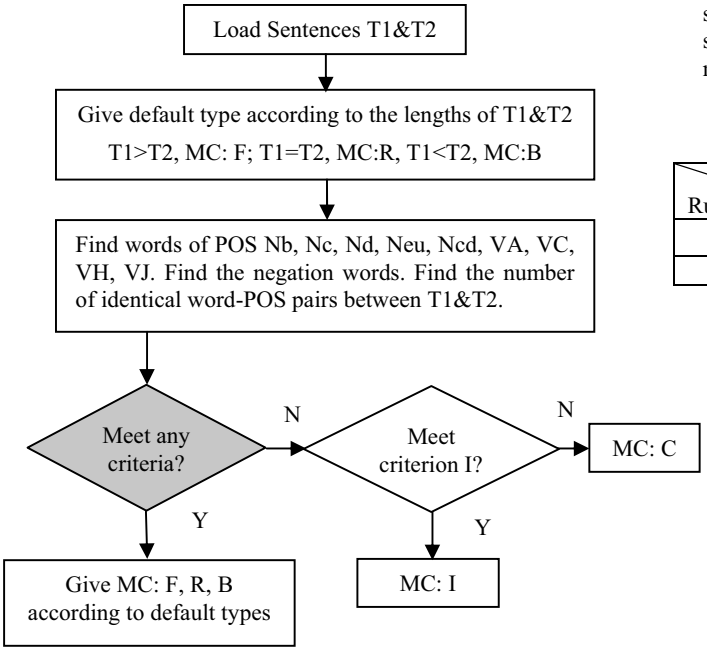

Figure 2. System flow for the MC subtask

3. Experiment

As the proposed system was rule based, the training sentence pairs were used only to do minor adjustments in the design of rules. Then these rules were applied directly on the testing sentence pairs. The proposed system was developed on traditional Chinese materials. Simplified Chinese sentences were translated into traditional Chinese ones before processing in our system.

3.1 Mapping Rules for BC in Run 2

As mentioned in section 2, we utilized the MC type results in Run 1 to generate BC results for Run 2 directly. The mapping rules are as follows:

1. Type F in MC Run 1 was mapped to type Y in BC Run 2.
2. Type B in MC Run 1 was mapped to type Y in BC Run 2.
3. Type R in MC Run 1 was mapped to type Y in BC Run 2.
4. Type C in MC Run 1 was mapped to type N in BC Run 2.
5. Type I in MC Run 1 was mapped to type N in BC Run 2.

Note that the third mapping rule was not a correct one. In fact, type R in MC should be mapped to type N in BC. For sentence pairs of type R, T1 cannot infer T2 by definition.

3.2 Experimental Results

Table 1 shows the experimental results of two runs. Run 1 achieved better results than Run 2. For the BC subtask, Run 1 performed better than Run 2 because one of the mapping rules was incorrect. If we fixed the rule, their results became comparable. For the MC subtask, Run 1 performed better than Run 2 suggested that giving more weights to the identical parts would not contribute to the performance, which responded to our assumption. Generally, the proposed preliminary system can compete some other systems in the traditional Chinese (TC) MC subtask, though the results were not good enough in other subtasks. In Table 2-5, the detail results from the confusion matrixes of Run 1 are shown further.

Table 1. Accuracy of Run 1 and Run 2 for the BC and MC subtasks

Accuracy / Run	BC (CS)	BC (CT)	MC (CS)	MC (CT)
1	0.636	0.528	0.528	0.477
2	0.560	0.524	0.398	0.388

Table 2. Confusion matrix of the BC subtask (simplified Chinese, run 1)

System / Answer	Y	N	
Y	243	128	371
N	20	16	36
	263	144	

Table 3. Confusion matrix of the BC subtask (traditional Chinese, run 1)

System / Answer	Y	N	
Y	404	379	783
N	46	71	117
	450	450	

Table 4. Confusion matrix of the MC subtask (simplified Chinese, run 1)

System \ Answer	F	R	B	C	I	
F	73	2	4	7	19	105
R	3	75	6	10	28	122
B	20	11	57	53	12	153
C	0	2	0	1	2	5
I	5	1	4	3	9	22
	101	91	71	74	70	

Table 5. Confusion matrix of the MC subtask (traditional Chinese, run 1)

System \ Answer	F	R	B	C	I	
F	122	6	25	29	48	230
R	6	136	16	28	49	235
B	42	24	126	105	40	337
C	0	4	3	5	3	15
I	10	10	10	13	40	83
	180	180	180	180	180	

Table 6. Precision for five types

Language \ Type	F	R	B	C	I
Simplified Chinese	72	82	80	1	12
Traditional Chinese	67	75	70	2	22

From the confusion matrix of the BC subtask, we found that our rules tended to judge sentence pairs as "able to infer" (Y, 128 pairs for SC and 379 pairs for TC). From the confusion matrix and the categorical precision of the MC subtask, we found that our rules had difficulty in identifying type C (precision 1%, 2%) and type I (precision 12%, 22%). We will discuss some possible causing issues in the next section.

4. Discussion

We checked the experimental materials according to the confusion matrix to find the problems of our system. Issues found are discussed in section 4.1 and section 4.2. The main problem of our system was that the rule to judge T1 and T2 as "able to infer" was too loose. The proposed system expanded the concept of relevance to determine the entailment and paraphrase, which was error prone because whenever one piece of additional or different information appeared in T2, T1 could not infer T2.

4.1 Error Analysis for the BC Subtask

As to the BC subtask, five issues were found:

1. The positions of identical words were different in two sentences, which made the meaning of two sentences different and the first sentence unable to infer the second one.

For example,

<pair id="549" label="N">

t1:日本松下電器指控 **台灣 (Taiwan)** 聯發科公司(MediaTek) 的晶片侵害 松下(Panasonic) 的專利

t2:日本松下電器指控 聯發科公司(MediaTek) 的晶片侵害**台灣 (Taiwan)** 松下(Panasonic) 的專利

<pair id="131" label="N">

t1:SARS **病毒 (SARS virus)** 屬於 (belongs to) **冠狀病毒 (coronavirus)**

t2:**冠狀病毒 (coronavirus)** 屬於 (belongs to) **SARS 病毒 (SARS virus)**

2. The different parts of two sentences were numbers or time expressed by Chinese numbers instead of Arabic numbers. A transformation component was needed to identify these numbers but we didn't have one in our system.

<pair id="799" label="N">

t1:陳盈豪在高中時智商測驗只有 **九十(90)** 左右

t2:陳盈豪在高中測驗時智商有 **一百二(120)** 左右

3. The units of the measurement were not converted to the same one and the comparisons were not processed by the system. Therefore, if they were different, the different proportion between two sentences was not large enough to judge it as unable to infer (label N).

For example,

<pair id="312" label="N">

t1:大陸「瞭望」雜誌的估算，東海油田蘊含石油 250 億 **公頓 (metric ton)**

t2:大陸「瞭望」雜誌的估算，東海油田蘊含石油 250 億 **噸 (ton)**

<pair id="174" label="N">

t1:英國經濟增長率 **大於(larger than)** 歐洲各國

t2:英國經濟增長率 **小於(smaller than)** 歐洲各國

4. The anaphors, co-references, and abbreviations were not processed by the system. It caused some false negatives.

For example,

<pair id="548" label="Y">

t1:日本松下電器指控台灣聯發科公司的晶片侵害 **松下 (Panasonic)** 的專利

t2:日本松下電器指控台灣聯發科技晶片侵害 **其(its)** 專利

<pair id="580" label="Y">

t1:邁可森將於 11 月 12 日搭乘 TG635 班機於晚間 7 點 40 分 **抵達 (arriving) 台灣 (Taiwan)**

t2:邁可森將於 11 月 12 日搭乘 TG635 班機於晚間 7 點 40 分 **抵台 (a short abbreviation for "arriving Taiwan" in news articles)**

5. The additional keywords did not provide additional information but restricted the meaning of modified words instead. Usually these keywords, referring to the different things in two sentences, are very similar, i.e., having the same subsequences. Words with underlines in the following examples show these subsequences.

For example,

<pair id="115" label="N">

t1:2004 年 **聽障 奧運 (Deaflympic)** 在雅典舉行

t2:2004 年 **奧運 (Olympic)** 在雅典舉行

<pair id="157" label="N">

t1:2005 年 7 月 7 日倫敦地鐵爆炸事件是在「**利物浦街站」(Liverpool St. Station)** 附近的地鐵「都會線」(Metropolitan line) 爆炸

t2:2005 年 7 月 7 日倫敦地鐵爆炸事件是在「**利物浦」(Liverpool)** 爆炸

4.2 Error Analysis for the MC Subtask

For the MC subtask, our system performed satisfactory for identifying forward entailment (F), paraphrase (B), and reverse entailment (R) types but badly for contradiction (C) and independence (I) types. The proposed system often mis-judged the contradiction type (C) as the paraphrase (B). This is because two sentences look very similar (usually only one difference) but:

1. The positions of identical words were different in two sentences, and there was no other different words or too few to determine them as a contradiction. This happened also in the BC subtask.

For example,

<pair id="565" label="C">

t1:20030501 **台北市立 (Taipei)** 和平醫院護理長陳靜秋因照顧 SARS 病患被感染於**林口 (Linkou)** 長庚醫院 (Chang Gung Hospital) 病逝 (passed away)

t2:20030501 **林口 (Linkou)** 和平醫院護理長陳靜秋因照顧 SARS 病患被感染於**台北市立 (Taipei)** 長庚醫院 (Chang Gung Hospital) 病逝 (passed away)

<pair id="786" label="C">

t1: 美國認為伊拉克具有 (has) **生化武器 (biochemical weapon)**，甚至 (even) 可能發展**核子武器 (nuclear weapon)**

t2:美國認為伊拉克具有 (has) **核子武器 (nuclear weapon)**，甚至 (even) 可能發展**生化武器 (biochemical weapon)**

2. No thesaurus or resources were involved in our system. Therefore, it had no ability to recognize synonyms (女兒, 千金, daughter) and antonyms (反感, disfavor, 好感, favorable) so that further judgments based on them were not feasible.

3. The processing of Chinese numbers and units of measurement was still an issue for the MC subtask. They needed to be converted into the same form.

It was difficult to find rules to identify the independence type (I) because various sentences could be of this type. Those sentence pairs of type I were wrongly judged as F, B and R types in our system, and this was because the decisions of this type were made mainly according to the sentence lengths since there was no other suitable rule to apply.

5. Conclusion and Future Work

We have designed a preliminary system 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. This simple rule-based system gave an overall passable performance and even a good performance for some types, but there was much room for improvement. Considering the order of words and modifying relations, involving thesaurus and resources, and converting numbers and units are some possible directions for instant improvements.

We found some important clues in the process of designing rules for our system. Therefore, beside the mentioned quick improvements, our next step is to utilize these clues by a learning process to implement a better rule-based, machine learning or hybrid system for determining types of sentence relations.

After the analysis, we also found that errors were due to two major phenomena: the alternation of the sentence meaning and the occurrence of the additional information. We will focus on observing the characteristics of these two phenomena to solve the system errors in the future and enhance the system performance.

6. ACKNOWLEDGMENTS

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC 100-2218-E-224-013-.

7. 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-9 Proceedings, to appear, 2011.

[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.