IASL RITE System at NTCIR-9

Cheng-Wei Shih, Cheng-Wei Lee, Ting-Hao Yang, Wen-Lian Hsu

Institute of Information Science, Academia Sinica, Taiwan, R.O.C

{dapi, aska, tinghaoyang, hsu}@iis.sinica.edu.tw

Abstract

We developed a knowledge-based textual inference recognition system for both BC and MC subtasks at NTCIR-9 RITE. Five different modules, which use named entities, subject-modifier word pairs, negative expressions, exclusive tokens and sentence length respectively, were implemented to determine the entailment relation of each sentence pair. Three decision making approaches were applied to integrate all the results from the recognition modules into one entailment result. The evaluation result showed that our system achieved 0.661 and 0.501 for traditional Chinese BC and MC subtasks respectively. For the simplified Chinese, the accuracy reached 0.715 and 0.565 for BC and MC respectively.

1. Introduction

Text understanding and inference, which is already believed as a necessary step in natural language application such as question answering, text summarization, and information retrieval, is one of the most challenging tasks in natural language processing. Therefore, determining the inference relation between two texts has become an important research topic since the First Recognizing Textual Entailment Challenge (RTE-1) hold in 2005 [1]. This year, NTCIR-9[2] provided a standard evaluation platform for Asia languages; aimed to help researchers focus on the text inference problem.

In RITE, All the systems were asked to classify the relations of sentence pairs (t1,t2) into both binary classes (Yes/No) and multiple classes (Forward, Reserve, Bidirectional, Contradiction, and Independent). Participants can freely use any language tools and knowledge resources to achieve the goal. We, team IASLD, aimed to recognize Chinese textual entailment relation in this task. The description of our work is organized as follows. Section 2 describes the system architecture. In Section 3, 4, and 5, we introduce the preprocess steps, the relation determining modules, and the decision making processing of entailment relation. Finally, we present the system performance in Section 6 and conclude our work in Section 7.

2. System Architecture

Our system focuses on knowledge-based approaches to classify five kinds of relations between two sentences. Several NLP tools and semantic resources are integrated into five different modules for relation recognition. We only aim at multiple-class classification. Our MC results are derived from the MC results. Figure 1 shows the structure of our system.

3. Preprocessing

In order to improve the accuracy of the output, some preprocessing steps are performed after the system receives each pairs. These steps include numerical character transformation and literal difference classification.

3.1. Numerical Character Transformation

All the numerical characters in numerical and temporal expressions are replaced by normalized digit forms. For example, a sentence such as “一九九九年十二月十日” (December 10, 1999) is converted to “1999 年 12 月 10 日” by substituting Chinese characters for digits. As not all sentences with Chinese numerical characters should be transformed, we need to be able to distinguish normal Chinese terms with numerical characters and numerical/temporal expressions. In our system, we use some hand-made regulations to target numerical and temporal expressions before the transformation.

On the other hand, in some numerical and temporal expressions such as range and duration, redundant parts are usually omitted. For example, a duration expression like “一九九九年十月至十二月”

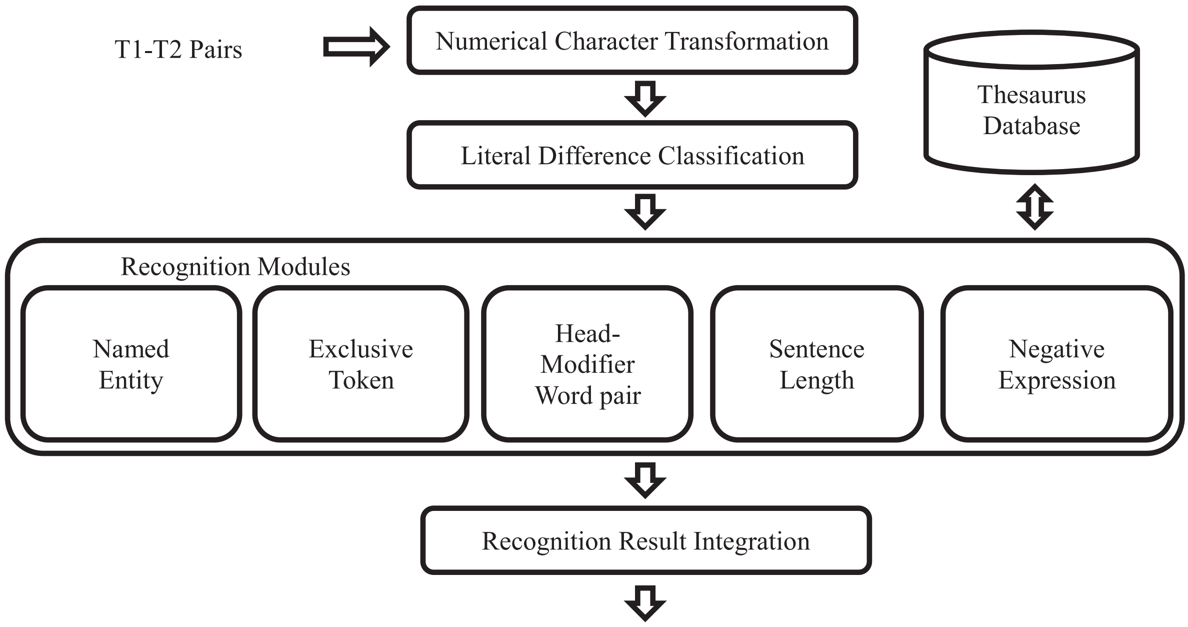

Figure 1. System Architecture.

(October to December, 1999) contains two time points. The year information “一九九九年” (1999) of the second time point does not explicitly expressed because it is the same as the first time point. Such an ellipsis may mislead the inference process and cause false recognition. We use several patterns to detect those omitted duration and range expressions and restore the ellipsis parts.

3.2. Literal Difference Classification

This module classifies a pair by the number of different sequences between t1 and t2. All the overlapping sequences between t1 and t2 are extracted. After that, the original t1 and t2 sentences are split into exclusive segments by these overlapping sequences. For example, given sentence t1:“思科公司是全球最大的網路供應公司” (Cisco is the biggest networking providing company) and t2:“微軟是全球最大軟體公司” (Microsoft is the biggest software company), the overlapping sequences are “是全球最大的” (…is the biggest…) and “公司” (company). We identify “思科公司” (Cisco) and “網路供應” (networking providing) as the exclusive segments of t1, “微軟” (Microsoft) and “軟體” (software) for the exclusive segments of t2. The rule we use to categorize the pairs is as follow :

We define C_{t1} as the number of exclusive segments in t1, C_{t2} as the number of exclusive segments in t2. The category of literal difference of t1 and t2:

$$LiteralDiffCategory(t1,t2) = \left\{ \begin{array}{l} "0:0", if\, C_{t1} = 0 \cap C_{t2} = 0 \\ "1:1", if\;\, C_{t1} = 1 \cap C_{t2} = 1 \\ "0:1", if\, C_{t1} = 0 \cap C_{t2} = 1 \\ "1:0", if\, C_{t1} = 1 \cap C_{t2} = 0 \\ "N:N", if\, C_{t1} > 1 \cap C_{t2} > 1 \\ "0:N", if\, C_{t1} = 0 \cap C_{t2} > 1 \\ "N:0", if\, C_{t1} > 1 \cap C_{t2} = 0 \end{array} \right\}$$

The type of literal difference is used to categorize the sentence pairs and determine a suitable recognition approach in the following process.

4. Entailment Recognition

Five recognition modules relying on different kinds of features commonly used in natural language processing are used independently to classify the relation of a pair. The relation categories, i.e. forward, reverse, bidirectional, contradictive, and independent, are identified by these five recognition modules. Besides, an extra “unknown” tag is provided for the case that a recognition module cannot classify the relation of a pair. These modules, which depend on named entities (NE), exclusive tokens, subject-modifier word pairs, sentence length, and negation expressions, will be introduced below.

4.1. Named-Entity-based Recognition Module

The main objective of this module is to detect all the exclusive named entities which only appear in one side of the pair and determine an entailment relation between the two sentences. Our named entity recognition tool, which was developed for CLQA and CCLQA in past NTCIR [3][4], integrates the results from a knowledge-based annotator and a CRF-based model. We use this NER tool to tag person names, location names, organization names, temporal expressions and numerical expressions in a sentence. All the common named entities of t1 and t2 are stripped away, and the entailment relation is determined by comparing the remaining NEs with pre-defined classifying rules.

This NE-based recognition module identifies forward, reverse, contradiction, and independence relation. However, for t1-t2 pairs with same set of NEs or no NE, NE-based module can only report "Unknown" as the result.

4.2. Exclusive-Token-Based Recognition Module

We make an assumption that the relation between two sentences can be decided by observing the relation of the exclusive segments of two sentences. In order to analyze the relations of exclusive segments of a pair, we use CKIP Chinese segmentation tool [5] to tokenize exclusive segments and get the exclusive tokens of t1 and t2 sentences. A mixture thesaurus of E-hownet [6], Chinese concept dictionary [7], Tongyichichilin [8], and Sinica Bow [9] is used to identify the semantic relations of all exclusive token pairs. Afterward, the following transforming rules are used to derive the entailment relations among exclusive tokens:

Assume t1 and t2 have exclusive token lists $ET_{t1} = \{et_{t1}1, et_{t1}2, \cdots et_{t1}N\}$ and $ET_{t2} = \{et_{t2}1, et_{t2}2, \cdots et_{t2}N\}$ respectively. The semantic relation labels between two words in the thesaurus are:

$$SR_{t1,t2} = \left\{ \begin{array}{l} Synonym, Hypernym, Hyponym, \\ Acronym, Meronym, Holonym, \\ Entailment, Cause, \ldots \end{array} \right\}$$

, then the inference direction between each two exclusive tokens $et_{t1}n$ and $et_{t2}m$ is defined as:

$$ID_{et_{t1}n, et_{t2}m} = \left\{ \begin{array}{c} Bidirectional, if\ SR_{et_{t1}n, et_{t2}m} = Synonym \\ Forward, if\ SR_{et_{t1}n, et_{t2}m} = Hypernym \mid Entailment \\ \mathrm{Re}verse, if\ SR_{et_{t1}n, et_{t2}m} = Hyponym \\ Contradiction, if\ SR_{et_{t1}n, et_{t2}m} = Acronym \\ Independent, otherwise \end{array} \right\}$$

The output of exclusive-token recognition can be defined as follow:

$$Output_{t,h} = \left\{ \begin{array}{l} Contradiction, if\ \exists ID_{et_{t1}n, et_{t2}m} = Contradiction \\ ArgMax(Num\_InferenceDirection_{\sum t1t \sum t2t}), \end{array} \right\}$$

where $t1t \in ET_{t1}, t2t \in ET_{t2}$.

In the use of the manually-defined rules, exclusive-token-based module can conclude the entailment relation from semantic relation of tokens. But relations such as independence cannot be reported for the lack of suitable matchup between semantic relations and entailment categories. For those pairs which cannot find any semantic relations, this module returns "unknown".

4.3. Head-Modifier-Word-Pair-Based Recognition Module

Compare to token-based analysis, we believe that the investigation of grammatically-related word pairs have a better chance to precisely retrieve useful information from a sentence. By using the structure of a parse tree, we can locate the linked word pairs and understand the modification correlation between two connected words in the parse tree.

In this recognition module, we use CKIP Chinese parser [10] to generate parse trees of t1 and t2. According to the part-of-speech tags and semantic-rule assignments provided by the parser, we can extract all the word pairs with head-modifier relations in the sentence. Figure 2 shows an example of extracting head-modifier word pairs from the sentence "廓爾喀族入侵尼泊爾" (Gurkhas invaded Nepal). In this example, the verb "入侵" (invaded) is the head of the sentence; the subject and object of the sentence associate with the head to become two head-modifier word pairs ("廓爾喀族-入侵" (Gurkhas invaded) and "入侵-尼泊爾" (invaded Nepal)) In order to reduce noisy information, a part-of-speech filter , which consists of five POS combination of head-modifier word pairs, is used to remove word pairs with stopwords and function words. Table 1 shows the POS combination restrictions and the examples.

Table 1. Allowable POS Combinations and the examples.

Adjective - Noun	重要-條件 (important - conditions)
Noun - Verb	廓爾喀族-入侵 (Gurkhas - invaded)
Verb - Noun	入侵-尼泊爾 (invaded - Nepal)
Adverb - Verb	成功-登上 (successfully - climb)
Noun - Noun	波斯灣-戰爭 (Gulf - War)

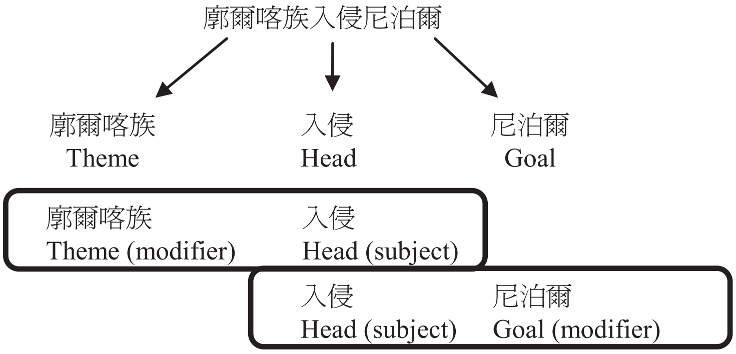

Figure 2. Example of extract Subject-modifier from parsing tree

A predefined rule is used to infer the semantic relation by checking the remaining head-modifier word pairs. The predefined rule is described below :

```
For each ( $TMS_i = \{m_i, s_i\}$  where  $m_i, s_i \in t1$ ) {
  For each ( $HMS_j = \{m_j, s_j\}$, where  $m_j, s_j \in t2$ )
  {  If( $ID_{s_i,s_j} = Bidirectional$ )
       The relation between $TMS_i$ and $HMS_j$
         $R\_MSWP_{TMS_i,HMS_j} = ID_{m_i,m_j}$
     Else if ( $ID_{m_i,m_j} = Bidirectional$ )
       The relation between $TMS_i$ and $HMS_j$
         $R\_MSWP_{TMS_i,HMS_j} = ID_{s_i,s_j}$
     Else
       The relation between $TMS_i$ and $HMS_j$
         $R\_MSWP_{TMS_i,HMS_j} = Independent$
     }
}
```

Head-modifier word pairs can help identify forward, reverse, contradiction, and independence relation. However, for those pairs with equal NEs or no NE, NE-based module reports an “Unknown” label.

4.4. Sentence Length-Based Recognition Module

Sentence length, in this task, is a useful feature to determine the direction of textual Entailment. Mostly, longer sentences own more information than short sentences. Using sentence length has a better chance to find correct results in forward and reverse cases, especially for pairs with larger length differences.

In our sentence length based module, literal length difference is not regarded as a feature directly. We use the number of tokens instead because we believe that the count difference of token can clearly indicates the distribution of information. The rule of sentence length based recognition is as follow: Assume L_{t1} and L_{t2} represent the token number of t1 and t2 respectively, then the output of sentence length module is:

$$Output_{SentenceLength}(t1,t2) = \left\{ \begin{array}{l} Forward, if\ L_{t1} > L_{t2} \\ \mathrm{Re}verse, if\ \ L_{t1} < L_{t2} \\ Bidirectional, if\ L_{t1} = L_{t2} \end{array} \right\}$$

As a result, sentence length based recognition module can only output forward, reverse and bidirectional relations.

4.5. Negative Expression Recognition Module

We think that most of the sentences with negation expression are negative sentences. Negation expression module, therefore, aim to capture negation expressions in a pair. In order to detect all possible negation expressions, we use a part-of-speech filter to extract negative terms from E-Hownet. If the exclusive token of each side of the pair contains negative expressions, we label it as contradiction.

As the rule is quite sample, negation expression module can only report if the pair relation is contradiction or not. For pairs without negative expressions, it outputs an “unknown” label.

5. Integration of Relation Recognition Results

We integrate the results from five recognition modules with two different methods: voting and predefined priority.

5.1. Voting

With a voting strategy, we choose the relation label that is supported by most of the recognition modules as the final result. For pairs that have more than one most supported labels, we generate a tie-break rule that can help us to determine which recognition module is supposed to be used for classifying the entailment relation.

We use two approaches to generate the tie-break rule. The first one is based on human experience. An analyst observed the development dataset, discovered the implicit rules or patterns, and arranged the priority of applying recognition module for different type of sentence pairs. The other approach is to find out the priority of module which is optimized to the development set. All the possible priority combinations of recognition module are generated and examined in the development sets. The priority combinations with best accuracy in the development sets will be adopted as the tie-break rule of the result integration.

5.2. Full-Manually Predefined Priorities

We decide the dominating recognition module based on full-manually predefined priorities. The entailment result from the designated module is directly treated as the final result. If the designated module outputs "unknown", the result of second high priority module is picked as the final result.

5.3. Confidence

The confidence value of a result r in our system is computed by the following formula:

$$Conidence_r = \frac{Num\,of\ Supporting\,Modules}{Total\ Num\ of\ Modules}$$

Theorically, a result with high confidence has a better change to be correct in the evaluation. We will examine the reliability of the confidence value afterward.

6. System Performance

There are BC and MC testsets for both simplified and traditional Chinese. Simplified and traditional Chinese test datasets contain 407 and 900 unlabeled t1-t2 pairs respectively.

Generally, our system was developed on the basis of using traditional Chinese knowledge sources and tools except the Chinese concept dictionary. For simplified Chinese pairs, we converted them into traditional Chinese pairs before the process. After the recognition process, the system's MC output is converted to BC labels for BC subtasks.

6.1. Official Evaluation Result

We submitted three different runs to compare the differences among the three integration strategies described in chapter 5. Table 2 shows the official evaluation results of our work.

Table 2. Evaluation result table of RITE

	CT-BC	CT-MC	CS-BC	CS-MC
Run1	0.648	0.499	**0.715**	**0.565**
Run2	0.653	0.487	0.705	0.543
Run3	**0.661**	**0.501**	0.688	0.555

6.2. Relation Recognition Analysis

In order to have a deeper insight of the system behavior and error trend, we compare the precision and recall of the 5 MC labels. The performance of the three runs of each subtask is shown in Table 3, 4, 5, and 6.

Table 3. System performances of multiple-class (MC) classification in traditional Chinese (CT)

	Run1	Run2	Run3
	Precision/Recall	Precision/Recall	Precision/Recall
F	0.548/0.728	0.559/0.733	0.575/0.617
R	0.558/0.722	0.518/0.711	0.580/0.706
B	0.478/0.361	0.463/0.344	0.487.0.317
C	0.368/0.333	0.369/0.367	0.388/0.317
I	0.488/0.35	0.481/0.278	0.442/0.55

Table 4. System performances of multiple-class (MC) classification in simplified Chinese (CS)

	Run1	Run2	Run3
	Precision/Recall	Precision/Recall	Precision/Recall
F	0.670/0.743	0.673/0.673	0.661/0.733
R	0.632/0.813	0.624/0.747	0.616/0.813
B	0.486/0.493	0.468/0.408	0.459/0.479
C	0.348/0.311	0.323/0.27	0.314/0.297
I	0.575/0.329	0.562/0.586	0.567/0.243

Table 5. System performances of binary-class (BC) classification in traditional Chinese (CT)

	Run1	Run2	Run3
	Precision/Recall	Precision/Recall	Precision/Recall
Y	0.634/0.698	0.644/0.684	0.674/0.622
N	0.664/0.598	0.664/0.622	0.649/0.7

Table 6. System performances of binary-class (BC) classification in simplified Chinese (CS)

	Run1	Run2	Run3
	Precision/Recall	Precision/Recall	Precision/Recall
Y	0.784/0.772	0.773/0.768	0.796/0.695
N	0.595/0.611	0.582/0.590	0.548/0.674

6.3. Recognition Module Analysis

Since our five recognition modules are independent, we can compute their MC accuracy on the test datasets separately. The accuracy results are demonstrated in Table 7 and 8. Because some recognition modules are capable to output "Unknown", the accuracy with/without counting unknown cases are also presented in the tables.

Table 7. The accuracy of five recognition modules in multiple-class classification in simplified Chinese. The rows in the table represent the performance of named entity based module, exclusive token-based module, head-modifier word pair module, sentence length-based module, and negative expression-based module respectively.

	with unknown	without unknown
NE	0.316	0.47
EToken	0.285	0.656
HMPair	0.333	0.471
SLength	0.465	0.465
NegativeE	0.056	0.432

Table 8. The accuracy of five recognition modules in multiple-class classification in traditional Chinese

	with unknown	without unknown
NE	0.3	0.425
EToken	0.226	0.547
HMPair	0.281	0.414
SLength	0.412	0.412
NegativeE	0.081	0.496

6.4. Confidence Analysis

In order to see the effectiveness of the confidence formula, we calculate the correlation between result accuracy and the confidence value. The correlation is depicted in figure 3 and 4.

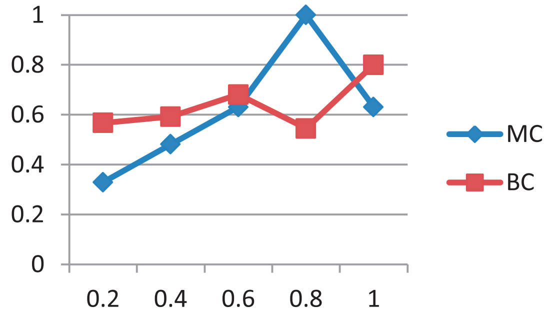

Figure 3. Correlation between accuracy and confidence of both binary-class and multiple-class recognition in traditional Chinese test set. The X-axis represents the confidence value; the Y-axis is the accuracy.

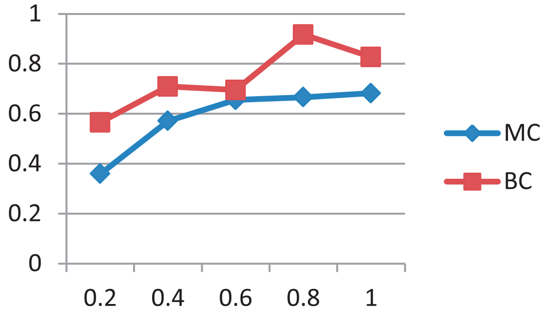

Figure 4. Correlation between accuracy and confidence of both binary-class and multiple-class recognition in simplified Chinese test set. The X-axis represents the confidence value; the Y-axis is the accuracy.

7. Conclusion

The system we developed for NTCIR-9 RITE integrates multiple knowledge-based recognition modules relying on shallow linguistic features such as Chinese tokens and named entities. The evaluation results show that the system performance is about the average among the participants in terms of identifying forward and reverse relations but unsatisfying in terms of the other three relation categories. This result may indicate several important things. First, for RITE datasets this year, sentence length is a useful but not reasonable feature which can effectively guide the entailment direction between two sentences. Second, the use of shallow linguistic features and hand-made thesaurus seems to be insufficient to distinguish equal, conflict, and unrelated sentences. However, even all of the modules, the idea of generating rules, and the module integration approaches that we described in this paper are very simple, our system is still a competitive baseline and a solid foundation for future researches of sentence entailment recognition.

Acknowledgments

This research was supported in part by the National Science Council under grant NSC 100-2631-S-001-001, and the research center for Humanities and Social Sciences under grant IIS-50-23.

Reference

[1] Ido Dagan, Oren Glickman and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F. (Eds.), *Machine Learning Challenges*. Lecture Notes in Computer Science, Vol. 3944, pp.

177-190, Springer, 2006.

[2] 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.

[3] Cheng-Wei Lee, Cheng-Wei Shih, Min-Yuh Day, Tzong-Han Tsai, Tian-Jian Jiang, Chia-Wei Wu, Cheng-Lung Sung, Yu-Ren Chen, Shih-Hung Wu and Wen-Lian Hsu (2005), "ASQA: Academia Sinica Question Answering System for NTCIR-5 CLQA," in Proceedings of NTCIR-5 Workshop, Tokyo, Japan, 202-208, (2005).

[4] Cheng-Wei Lee, Min-Yuh Day, Cheng-Lung Sung, Yi-Hsun Lee, Tian-Jian Jiang, Chia-Wei Wu, Cheng-Wei Shih, Yu-Ren Chen, Wen-Lian Hsu, "Chinese-Chinese and English-Chinese Question Answering with ASQA at NTCIR-6 CLQA", in Proceedings of NTCIR-6 Workshop Meeting, Tokyo, Japan, May 15-18, 2007, pp. 175-181. (2007).

[5] Ma, Wei-Yun and Keh-Jiann Chen, 2003, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp168-171.

[6] Chen Keh-Jiann, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen, 2005, Extended-HowNet- A Representational Framework for Concepts, OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop, Jeju Island, South Korea.

[7] Yu J.S. and Yu S.W. et al 2001 Introduction to Chinese Concept Dictionary, in International Conference on Chinese Computing (ICCC2001), pp361-367.

[8] 梅家驹,竺一鸣，高蕴琦等编.同义词词林.上海：上海辞书出版社，1983

[9] Chu-Ren Huang, Ru-Yng Chang, and Shiang-Bin Lee. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. In Proceedings of LREC2004, pages 1553- 1556, Lisbon, Portugal.

[10] Chinese Parser, http://ckip.iis.sinica.edu.tw/CKIP/parser.htm