Description of the NTOU RITE-VAL System at NTCIR-11

Chuan-Jie Lin, Chi-Ting Liu, and Yu-Cheng Tu
Department of Computer Science and Engineering
National Taiwan Ocean University
2 Pei-Ning Road, Keelung, Taiwan 20224
{cjlin,ct.liu,yctu.cse}@ntou.edu.tw

ABSTRACT

System validation subtask in NTCIR aims at developing techniques to deal with many kinds of language phenomena about textual entailment. This paper introduces our system participating in NTCIR-11 RITE-VAL SV Subtask. By adopting different combination of features related to WordNet, Tongyici Cilin, and syntactic information, 5 SV-BC and 5 SV-MC formal runs were submitted. The best BC run achieved 42.89% in macro F-measure and 52.33% in accuracy. The best MC run achieved 31.03% in macro F-measure and 39.17% in accuracy.

Team Name

NTOUA

Subtasks

System Validation (Traditional Chinese)

Keywords

textual entailment, system validation, WordNet, Tongyici Cilin, machine learning

1. INTRODUCTION

The NTOUA team participated in the system validation (SV) subtask of the NTCIR-11 RITE-VAL task [11] this year. Recognizing textual entailment (RTE) has been studied for several years, such as in the TAC RTE tracks [1] and EVALITA IRTE task [2]. The RTE techniques are useful in many research areas, such as answer validation in question answering [12] and text extraction in summarization [9]. It is our third attempt to develop a Chinese RTE system [6, 7]. This paper describes our approaches to solve the RTE problem and reports the official results.

There are two types of SV subtask: binary-class (BC) and multi-class (MC). Given a pair of sentences (t_1, t_2), the BC subtask is to determine whether t_1 entails t_2, while MC subtask is to determine the entailment direction or contradiction. The labels used in BC subtask are "Y" and "N." The labels defined in MC subtask are "F" (for forward entailment, $t_1 \Rightarrow t_2$), "B" (for bidirectional entailment, $t_1 \Leftrightarrow t_2$), "C" (for contradiction), and "I" (for independence). We participated in both subtasks.

To tackle various kinds of language phenomena related to textual entailment, we proposed 46 features, including WordNet semantic features, Cilin semantic features, syntactic features and lexical features. Systems were SVM classifiers trained by using different combination of features.

This paper is organized as follows. Section 2 describes text preprocessing and feature definition. Section 3 explains the development of classifiers and different system settings. Section 4 shows the evaluation results of formal runs and Section 5 gives conclusions.

2. FEATURE DEFINITION

Our system is mainly SVM classifiers trained by using several features concerning surface and sense similarities. We tested on different feature settings and submitted five formal runs in each subtask to see the applicability of the proposed features. Proposed features and RITE systems are explained in this and next sections.

Text preprocessing on sentences in the training sets and test sets includes Chinese word segmentation, part-of-speech (POS) tagging, syntax, parsing, named-entity recognition (NER), temporal and numerical information resolution. All systems were built in our lab except Stanford dependency parser [4, 10].

Based on the characteristics of Chinese POS, only normal nouns, proper nouns, and verbs were considered as content words in our experiment. The information of person, location, and time is important when describing an event. Therefore, person names and location names were identified by our NER system. Date and numerical expressions were extracted by patterns. Moreover, temporal and numerical information would be transformed into canonical forms if possible.

In order to catch more contemporary terms, we also extracted occurrences of Wikipedia titles in the sentence pair. We adopted the matching method proposed by Lin and Tu [8]. Note that the method is based on longest matching strategy and the boundaries of Wikipedia titles would take over the ones by word segmentation.

For each sentence t_i, $i \in (1, 2)$, the following sets were created for sentence comparison and similarity scoring:

- W_i, the set of distinct content words in t_i
- O, the set of overlapped words, i.e. $W_1 \cap W_2$
- D_i, the set of different words in t_i, i.e. $W_i - O$

From our observation, words in the sentences in a text pair are almost identical except one or two of them are different. These different words determine entailment or contradiction. So we proposed several features mainly concerning these different words.

On the other hand, if two sentences have identical word sets, syntax instead determines entailment or contradiction. Hence more features related to syntax are also proposed.

46 features were used to train our classifiers. Feature values were determined by WordNet, Tongyici Cilin, and the dependency relations from Stanford dependency parser. The definitions of the features are as follows.

2.1 WordNet Semantic Features

WordNet semantic features are used to capture word sense similarity and difference between two sentences by using WordNet. Besides sense similarity, hypernymy and antonymy is also important in determining entailment and contradiction.

WordNet is a thesaurus of English words, while Sinica BOW[1] [5] provides Chinese translations of synsets in WordNet. We used these two dictionaries together to measure sense similarities of Chinese words.

Some functions needed for feature extraction are defined as follows.

- $\text{score}_{WN+}(w_i, w_j)$. Given two words w_i and w_j, their senses in WordNet are first located. Let ca be their nearest common ancestor in hypernymy, d_i and d_j the lengths of paths from them to their common ancestor ca, and dc the length of the path from the root to ca. If more than one possible path is found, the shortest one is selected. We adopt Wu and Palmer's [13] definition of WordNet similarity defined as:

$$\text{score}_{WN+}(w_i, w_j) = \frac{2 \times dc}{(dc + d_i) + (dc + d_j)}$$

  When two words do not have a common ancestor in WordNet, their similarity score is defined as 0. When a word has more than one sense or a sense has more than one hypernym, the largest score among all cases is chosen.

- $\text{score}_{WN-}(w_i, w_j)$. The WordNet relatedness score in antonymy relationship is defined in the similar way except that a pair of senses (ca_i, ca_j) is to be found where ca_i is ancestor of w_i and ca_j is ancestor of w_j in hympernym, and ca_i is an antonym to ca_j.

- $\text{isHyp}_{WN}(w_i, w_j) = 1$ if w_j is a hypernym of w_i, i.e. $\text{score}_{WN+}(w_i, w_j) > 0$ and $d_j = 0$; 0 otherwise.

- $\text{isAnt}_{WN}(w_i, w_j) = 1$ if w_i and w_j are antonyms, i.e. $\text{score}_{WN-}(w_i, w_j) > 0$, $d_i = 0$ and $d_j = 0$; 0 otherwise.

- $\text{isSim4}_{WN}(w_i, w_j) = 1$ if w_i and w_j have similar senses, which is defined as $\text{score}_{WN+}(w_i, w_j) > 0$, $d_i \neq 0$, $d_j \neq 0$, and $d_i + d_j \leq 4$; 0 otherwise.

20 WordNet semantic features are defined as follows.

- 4 features of $\sum \text{score}_{WN+}(w_i, w_j)$ over all (w_i, w_j) pairs
- 4 features of $\sum \text{score}_{WN-}(w_i, w_j)$ over all (w_i, w_j) pairs
- 4 features of $\bigvee \text{isHyp}_{WN}(w_i, w_j)$ over all (w_i, w_j) pairs
- 4 features of $\bigvee \text{isAnt}_{WN}(w_i, w_j)$ over all (w_i, w_j) pairs
- 4 features of $\bigvee \text{isSim4}_{WN}(w_i, w_j)$ over all (w_i, w_j) pairs

where $(w_i, w_j) \in D_1 \times D_2$, $D_1 \times O$, $D_2 \times D_1$, or $D_2 \times O$, respectively.

[1] `http://bow.ling.sinica.edu.tw/`

2.2 Cilin Semantic Features

Tongyici Cilin is a thesaurus of Chinese words, which collects more Chinese words than Sinica BOW. A project held in Harbin Institute of Technology further expanded it with more modern words. Now the extended version contains 17,817 synsets and 77,371 distinct Chinese words. We used the HIT extended edition Cilin to measure word sense similarity.

Cilin semantic features are used to capture word sense similarity and difference between two sentences by using Cilin. Unlike WordNet, Cilin does not record hypernymy relationships. We will measure sense similarity by Cilin synset IDs.

Cilin assigns each synset with a unique ID with a format of $XymZn\$$, where X and Z are uppercase letters, y is a lowercase letter, m and n are two-digit numbers, and a symbol \$. Cilin organizes senses in a 5-layer hierarchy, which will be represented by the first five codes in a synset ID. We will refer to a specific code by saying "the k^{th}-level ID code" of a synset hereafter in this paper. Take the synset ID "Bc03A01=" as an example, its 5^{th}-level ID code is "01".

If two Cilin synset IDs share longer common heading strings, the words in these two synsets have more similar meanings. Therefore, we measure sense similarity by the surface similarity of synset IDs.

The trailing symbol \$ in a synset ID is one of $\{=, \#, @\}$, where '=' means a general synset, '#' denotes members belonging to a set but not synonyms (e.g. names of holidays), and '@' means that the synset contains only one word.

Some functions needed for feature extraction are defined as follows.

- $comLevel(y_i, y_j)$ is number of the common leading level codes in synsets IDs of y_i and y_j. For example, $comLevel(Bc03A01, Bc03A02) = 4$ and $comLevel(Ac03A02, Bc03A02) = 0$.

- $\text{score}_{Cilin}(w_i, w_j)$ is defined as

$$\text{score}_{Cilin}(w_i, w_j) = \frac{comLevel(y_i, y_j)}{5}$$

  where w_i belongs to the synset y_i and w_j belongs to the synset y_j.

  If $comLevel(y_i, y_j) = 5$ and the last code of ID is '#', $\text{score}_{Cilin}(w_i, w_j)$ is defined as 0.8 because they are not real synonyms.

  If any word belongs to more than one synset, the highest score_{Cilin} among the synset pairs is chosen.

- $\text{isHyp}_{Cilin}(w_i, w_j) = 1$ if w_j is a hypernym of w_i; 0 otherwise.

  We find that the first synset (i.e. $n = 01$) in a 5^{th}-level class is often a hypernym to the other synsets in the same class (i.e. same $XymZ$ but $n \neq 01$).

  Therefore, $\text{isHyp}_{Cilin}(w_i, w_j) = 1$ if and only if $comLevel(y_i, y_j) = 4$ and the 5^{th}-level ID code of y_j is "01".

- $\text{isSim3}_{Cilin}(w_i, w_j) = 1$ if w_i and w_j have similar senses, which is defined as $3 \leq comLevel(y_i, y_j) \leq 4$ and $\text{isHyp}_{Cilin}(w_i, w_j) = 0$; 0 otherwise.

- $\text{isMem}_{Cilin}(w_i, w_j) = 1$ if w_i and w_j are members in the same set, which is defined as $comLevel(y_i, y_j) = 5$ and the trailing symbol of ID is '#'; 0 otherwise.

16 Cilin semantic features are defined as follows.

- 4 features of $\sum \text{score}_{Cilin}(w_i, w_j)$ over all (w_i, w_j) pairs
- 4 features of $\bigvee \text{isHyp}_{Cilin}(w_i, w_j)$ over all (w_i, w_j) pairs
- 4 features of $\bigvee \text{isSim3}_{Cilin}(w_i, w_j)$ over all (w_i, w_j) pairs
- 4 features of $\bigvee \text{isMem}_{Cilin}(w_i, w_j)$ over all (w_i, w_j) pairs

where $(w_i, w_j) \in D_1 \times D_2$, $D_1 \times O$, $D_2 \times D_1$, or $D_2 \times O$, respectively.

2.3 Syntactical Features

Two sentences may be contradictory if they have the same subjects and verbs but different objects, or the subject and the object are exchanged. In order to capture syntactic information, we used Stanford dependency parser to build dependency trees.

Dependency relations in a dependency tree are in the format of (a_i, dep, a_j), which means there is a dependency relationship dep between two arguments, words a_i and a_j.

2 syntactic features are defined as follows.

- $arg_{diff} = 1$ if there exists two dependency relations having the same dependency relationship and one argument but the other argument is different; 0 otherwise.
- $dep_{diff} = 1$ if there exists two dependency relations having same arguments but different dependency relationship; 0 otherwise.

We only consider dependency relationship in $\{mod, obj, subj\}$, i.e. modifiers, objects, and subjects.

2.4 Lexical Features

Concerning content words, their intersection and different sets, 8 features are proposed to capture surface and lexical information. Note that words were lemmatized at the beginning.

- 4 features capture the degree of overlap between two sentences. These features are defined by a synonym set V and the function $\text{overlap}(S_i, S_j)$ defined as:

$$\text{overlap}(S_i, S_j) = \frac{|V|}{|S_i|}.$$

The synonym set V is defined as follows. Given two non-overlapping word sets, for each $w_i \in S_i$ and $w_j \in S_j$, if one of the following conditions is true then $V \leftarrow V \cup \{w_i\}$.

1. w_i and w_j are synonyms in WordNet
2. w_j is a hypernym of w_i in WordNet
3. w_i and w_j are synonyms in Cilin
4. w_j is a hypernym of w_i in Cilin, i.e. $\text{isHyp}_{Cilin}(w_i, w_j) = 1$.

If S_i is an empty set, $\text{overlap}(S_i, S_j) = 0$.

These 4 features related to synonymy overlapping are

  - $\text{overlap}(D_1, W_2)$
  - $\text{overlap}(D_2, W_1)$
  - $1 - \text{overlap}(D_1, W_2)$; defined as 0 if $D_1 = \varnothing$
  - $1 - \text{overlap}(D_2, W_1)$; defined as 0 if $D_2 = \varnothing$

- 1 feature denotes whether words in t_2 are substrings of t_1. I.e. $\text{hasSubstr}(W_1, W_2) = 1$ if there are $w_1 \in W_1$ and $w_2 \in W_2$ such that w_2 is a substring of w_1 and $w_2 \neq w_1$; 0 otherwise.
- 2 features compare sizes between the difference sets. They are $|D_1| - |D_2|$ and $|D_2| - |D_1|$.
- 1 feature shows that two sentences have identical word sets. I.e. $\text{ident}_{\text{word}}(W_1, W_2) = 1$ if $D_1 = \varnothing$ and $D_2 = \varnothing$; 0 otherwise.

3. NTOU RITE-VAL SYSTEMS

In this RITE-VAL Task, we built 5 systems to do the experiments. Four of them are multi-class classifiers using different sets of features. Their feature settings are as follows.

- SYS_1 uses all features
- SYS_2 uses all but WordNet semantic features
- SYS_3 uses all but Cilin semantic features
- SYS_4 uses all but syntactic features

The fifth system SYS_5 consists of 3 binary classifiers (denoted as S_{YN}, S_{FB}, and S_{CI}) in a 2-layer hierarchy. A sentence pair is first classified by S_{YN} which determines where t_1 entails t_2. If the sentence pair is classified as 'Y' (entailment), it is further classified by S_{FB} to determine if the entailment is bi-directional. On the other hand, if this sentence pair is classified as 'N' (non-entailment), it is further classified by S_{CI} to determine if the sentences are contradictory.

We used CT-SVMC pairs labeled with F and B to train S_{FB} and CT-SVMC pairs labeled with C and I to train S_{CI}, but CT-SVBC training set to train S_{YN}.

All 7 classifiers were trained by LIBSVM [3]. We also used its tools to find the best parameter settings on training data.

4. EVALUATION RESULTS

Five CT-SVMC formal runs output by five systems were submitted to the CT-SVMC subtask.

To produce CT-SVBC formal runs, the CT-SVBC test set was first predicted by these 5 multi-class systems. By changing labels F and B into label Y (entailment), and labels C and I into label N, 5 corresponding SVBC were created and submitted to the CT-SVBC subtask.

The formal evaluation metric of BC and MC subtasks are macro F-measure (MacroF1, the average of F-measures of every labels) and accuracy score (Acc., the ratio of correctly predicted pairs). The run scores marked in **bold** in the tables are the best results. Tables 1 and 2 show the evaluation results of all the runs in the SV-BC and SV-MC subtasks, respectively.

In the five official runs of each subtask, the third system (a classifier trained without Chinese WordNet [Sinica BOW]) achieves the best performance in both macro F-measure and

Run	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
NTOUA-CT-SVBC-01	39.26	50.67	65.58	50.36	94.00	12.94	55.00	7.33
NTOUA-CT-SVBC-02	39.26	50.67	65.58	50.36	94.00	12.94	55.00	7.33
NTOUA-CT-SVBC-03	**42.89**	**52.33**	66.11	**51.29**	93.00	**19.66**	62.50	**11.67**
NTOUA-CT-SVBC-04	41.01	49.83	63.82	49.91	88.50	18.21	49.26	11.17
NTOUA-CT-SVBC-05	39.73	51.67	**66.55**	50.88	**96.17**	12.91	**65.15**	7.17

Table 1: Performance of the official runs on SV-BC subtask.

Run	MacroF1	Acc.	B-F1	B-Prec.	B-Rec.	F-F1	F-Prec.	F-Rec.	C-F1	C-Prec.	C-Rec.	I-F1	I-Prec.	I-Rec.
NTOUA-CT-SVMC-01	28.89	38.33	48.68	**35.59**	77.00	53.44	43.74	68.67	10.47	40.91	6.00	2.98	13.89	1.67
NTOUA-CT-SVMC-02	28.83	38.25	48.47	35.50	76.33	53.42	43.58	**69.00**	10.47	40.91	6.00	2.98	13.89	1.67
NTOUA-CT-SVMC-03	**31.03**	**39.17**	**49.45**	35.22	**83.00**	52.86	**47.24**	60.00	**16.79**	35.48	**11.00**	5.02	**42.11**	2.67
NTOUA-CT-SVMC-04	29.31	37.17	49.35	35.18	82.67	50.08	45.96	55.00	10.47	24.39	6.67	**7.34**	24.07	**4.33**
NTOUA-CT-SVMC-05	29.03	38.58	48.45	34.57	81.00	**53.90**	45.71	65.67	9.50	**43.24**	5.33	4.26	24.14	2.33

Table 2: Performance of the official runs on SV-MC subtask.

accuracy. Two systems utilizing all features (f_{all}, an MC classifier, and fifth system, a two-stage model using three BC classifiers) are worse than the two single-semantic-resource systems. However, the difference between feature selections is not obvious and stable. There is no conclusion which selection is better than the other.

5. CONCLUSION

This paper described the approaches of our system to recognize semantic relations between sentences in the NTCIR-11 RITE-VAL task. Several features have been proposed to train entailment relationship classifiers and total 10 formal runs were submitted. Again, our best system in the BC subtask achieves 42.89% in macro F-measure and 52.33% in accuracy. The performance of our MC classifiers is around 31.03% in macro F-measure and 39.17% in accuracy.

We adopted different sets of features which were extracted from many well-known resources to do machine learning. In the future, we plan to study the efficiency of each feature and find out the best combination. The advantages and weaknesses of our system will also be observed under the characteristics of NTCIR RITE-VAL training datasets.

6. REFERENCES

[1] L. Bentivogli, P. Clark, I. Dagan, H. T. Dang, and D. Giampiccolo. The Sixth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Third Text Analysis Conference (TAC)*, Gaithersburg, MD, USA, 2010. National Institute of Standards and Technology.

[2] J. Bos, F. M. Zanzotto, and M. Pennacchiotti. Textual Entailment at EVALITA 2009. In *Proceedings of the Second Workshop on Evaluation of NLP and Speech Tools for Italian (EVALITA)*, Reggio Emilia, Italy, 2009. Italian Association of Speech Science.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] P.-C. Chang, H. Tseng, D. Jurafsky, and C. D. Manning. Discriminative Reordering with Chinese Grammatical Relations Features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST '09)*, pages 51–59, 2009.

[5] C.-R. Huang. Sinica BOW: integrating bilingual WordNet and SUMO ontology. In *Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering*, pages 825–826, 2003.

[6] C.-J. Lin and B.-Y. Hsiao. The Description of the NTOU RITE System in NTCIR-9. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 353–356, Tokyo, Japan, 2011.

[7] C.-J. Lin and Y.-C. Tu. The Description of the NTOU RITE System in NTCIR-10. In *Proceedings of the 10th NTCIR Conference*, pages 495–498, Tokyo, Japan, 2013.

[8] C.-J. Lin and Y.-C. Tu. Word Segmentation Refinement by Wikipedia for Textual Entailment. In *Proceedings of the 15th IEEE International Conference on Information Reuse and Integration, Workshop on Empirical Methods for Recognizing Inference in TExt (IEEE EM-RITE 2014)*, pages 600–606, 2014.

[9] E. Lloret, O. Ferrández, R. Muñoz, and M. Palomar. A Text Summarization Approach under the Influence of Textual Entailment. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS)*, pages 22–31, Barcelona, Spain, 2008.

[10] M.-C. d. Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from

phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 449–454, 2006.

[11] S. Matsuyoshi, Y. Watanabe, Y. Miyao, T. Shibata, T. Mitamura, C.-J. Lin, and C.-W. Shih. Overview of the Recognizing Inference in Text and Validation (RITE-VAL) at NTCIR-11. In *Proceedings of the 11th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-11)*, Tokyo, Japan, 2014. to appear.

[12] A. Rodrigo, A. Peñas, and F. Verdejo. Overview of the Answer Validation Exercise 2008. In *Proceedings of the 9th Workshop on Cross-Language Evaluation Forum (CLEF)*, volume 5706 of *Lecture Notes in Computer Science*, pages 296–313, Aarhus, Denmark, 2008.

[13] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, 1994.