Combining Multiple Lexical Resources for Chinese Textual Entailment Recognition

Yu-Chieh Wu	Yue-Shi Lee	Jie-Chi Yang
Dept. of Communication and Management	*Department of Computer Sciences and Information Engineering*	*Graduate Institute of Network Learning Technology*
Ming-Chuan University	*Ming-Chuan University*	*National Central University*
Taipei, Taiwan	*Taipei, Taiwan*	*Jong-Li City, Taiwan*
wuyc@mail.mcu.edu.tw	*leeys@mail.mcu.edu.tw*	*yang@cl.ncu.edu.tw*

Abstract

Identifying Textual Entailment is the task of finding the relationship between the given hypothesis and text fragments. Developing a high-performance text paraphrasing system usually requires rich external knowledge such as syntactic parsing, thesaurus which is limited in Chinese since the Chinese word segmentation problem should be resolve first. By following last year, in this year, we continue adopting the created RITE system and combine with multiple online available thesaurus. We derive two exclusive feature sets for learners. One is the operations between the text pairs, while the other adopted the traditional bag-of-words model. Finally, we train the classifier with the above features. The official results indicate the effectiveness of our method.

1. Introduction

Discover textual relations using paraphrasing techniques shows successful results in recent years. Textual information which is one of the most important features for most text mining research issues

The textual entailment recognition task aims to identify, given two text snippets *t* and *h*, whether *t* entails *h* or not (where *t* means the entailing text and *h* is the hypothesis or the entailed text). The goal of NTCIR-RITE challenge (Watanabe et al., 2013) has been to create traditional and simplified Chinese and Japanese benchmark corpus dedicated to textual entailment – recognizing that the meaning of one text is entailed. This task is very competitive and raised many text mining techniques, such as Natural Language Processing (NLP) (Manning and Schutze, 1999), Information Extraction (IE), Chinese Text Processing (CTP), Machine Learning (ML) etc. Textual entailment (aka paraphrasing) provides useful information for downstream purposes. Examples include, question answering (Voorhees, 2001; Oh et al., 2007), sentence compression, text summarization, and sentence rephrasing.

English textual entailment has been addressed well in past few years. The well-known PASCAL workshop on RTE (Giampiccolo et al., 2007) is the best example. NTCIR RITE opens a very early competition on the task of Asian text entailment. It comes up with four different languages, English, Japanese, and (simplified and traditional) Chinese. Participants have to choose BC (binary) or MC (multiple) or QA (question answering) or partial of them and submit the result. BC is simply to identify the entailment relation between the given pair of text fragments (yes or no), while MC is to label the relation of the given fragment. In this year, we only focus on the BC and MC tasks for traditional Chinese.

Chinese textual entailment is a new open research issue. There are fewer literatures about

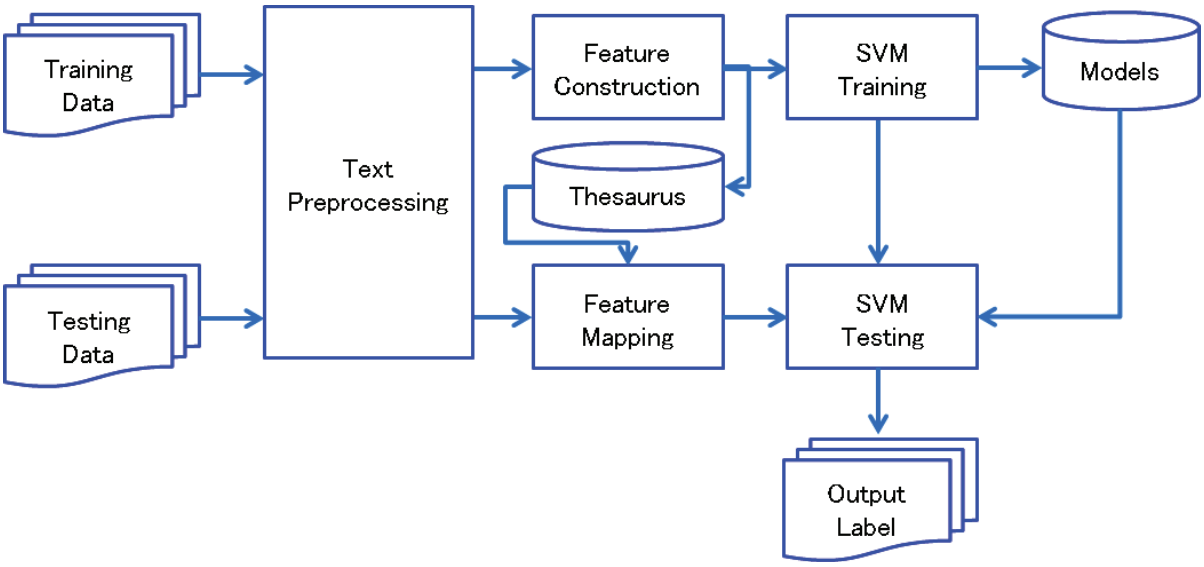

Figure 1: Overall System Flowchart

this topic. Huang et al. (2011) presented a complex Chinese textual entailment recognition system. Due to the lack of traditional Chinese syntactic parser, they convert the text into simplified Chinese for parsing. Furthermore, they propose many heuristics to correct the Chinese word segmentation errors and numeric text normalization. They employed the LibSVM (Lin et al., 2005) to learn to find the textual entailment relation. As reported by (Huang et al., 2011), the most useful feature is the "tree mapping" which requires a parser. In English textual entailment (Androutsopoulos and Malakasiotis, 2010), a set of approaches were proposed. For example, the logic proofer (Tatu and Moldovan, 2007), machine learning-based (Li et al., 2007; Malakasiotis, 2009), similarity-based (Malakasiotis and Androutsopoulos, 2007; Wang and Neumann, 2007), syntactic similarity-based (Wan et al., 2006) and hybrid approaches (Tatu and Moldovan, 2007). However, those methods are quite difficult to port to Chinese. The biggest challenge is that there is no explicit word boundary between words. Also, the resources (like parser, thesaurus) for Chinese is limited.

In this paper, we follow the previous developed RITE system (Wu et al., 2011) by extending the lexical information. Similar to our previous system, both Chinese word segmentation and POS tagging showed powerful results in recognition. The derived online lexicons were used to improve the overall lexical coverage. Our method combines both statistical and lexical features. We propose a set of features for learners. Some of features are shallow syntactic pattern-based with only POS tag information while some of them are estimated from the training data.

The remainder of this paper is organized as follows: Section 2 gives the analysis of RITE task data set used in this year. Section 3 introduces title proposed method. Section 4 discusses the experiment settings and the performance results in this competition. Finally, concluding remarks are drew in section 5.

2. System Overview

Figure 1 shows the proposed RITE system used in the NTCIR-RITE this year. The first component (text preprocessing) is to firstly segment Chinese text and give POS tag for each word. Second, we construct the feature set (plus the employed external thesaurus) from the training data and mapping training instance for SVM. In the third stage, the SVM training and testing modules receive the instances and performing learning and classifying. The trained SVM model is used to forecast the testing data and give the entailment label for each pair of text fragment. In the following section, we introduce the three important modules.

3. Approach

3.1 Preprocessing

Text mining in Chinese is quite difficult than most western languages, such as English due to the word information is not available in text. There is no explicit word boundary between words in Chinese text. To resolve this, a Chinese word segmentation tool is needed. It plays an important role the preprocessing step since word information provides the basic concept in term-level for downstream applications. In addition, the POS tag information also gives basic syntactic structures in text. However, there are few Chinese word segmentation tools for our purpose, in this paper, we revise our in-house CMM-based Chinese word segmentation and POS tagging method (Wu et al., 2009, 2010a, 2010b).

3.2 Text Normalization

A set of Chinese words share the same meanings as Arabic numbers, such as 伍 equals 5. Also the holomorphy words need to be normalized. However, directly transform these words into numbers is not a good idea, some words might be partial of a person name. To solve this, the normalization process only deals with a small set of POS tags. For Neu (number) and Nd (date) words, Chinese numerical words are directly converted to digits. For example, 亅-> 1, 二->2, 叁-> 3, etc.

There are still some complex Chinese words express numbers, such as 二十一- > 21. A simple rule is designed to solve this. If a specified Chinese word is find (十、廿、卅、百、千、萬), a left-right search is also applied. For all Chinese numeric words that locates on the left hand-side of the specified word, the numeric words were converted using the above text normalization method and multiply the specified word. Similarly, for all the right hand side Chinese numeric words were normalized and plus the left hand side numbers.

3.3 Feature Construction

For better prediction power, we construct two different feature types, namely statistical features and lexical features. The former measures the general statistic information of each paired texts, while the later captures the lexical-level information using both Chinese word and POS tags. Below, we list the used features in this paper.

Type I

- Length difference (character-level)
- Length difference (word-level)
- Character match ratio in s_1
- Character match ratio in s_2
- Word match ratio in s_1
- Word match ratio in s_2
- POS match ratio in s_1
- POS match ratio in s_2
- Pattern match ratio in s_2
- Pattern match ratio in s_2
- Reversed pattern match ratio in s_2
- Reversed pattern match ratio in s_2
- Minimum number difference

Type II

- Matched POS tags
- Matched Bi-POS tags
- Mismatched POS tags
- Matched Verb tags
- Mismatched Verb tags
- Mismatched Verb words

Here, the pattern is predefined as the specified POS bigram and trigrams. We define the following six patterns.

Noun+Verb, Verb+Noun, Noun+Noun,
Noun+Verb+Noun, Verb+Noun+Verb,
Noun+Noun+Noun

Even the six patterns are defined to find the matched statistics. We also reverse the *order* for each pattern. That is, the reversed patterns can be used to find the contradiction sentence pairs. To enhance the results, both word and POS tag were used to represent the pattern. For example, the first pattern, Noun+Verb, the word bigram and POS bigram were extracted. In total, there 6*2(POS and Word)*2(plus reverse order) = 24 patterns were extracted.

3.4 Thesaurus Feature

In this year, we further added thesaurus information to enhance the lexical coverage, Three different types of the lexicons were adopted, namely, Ciling, positive/negative word set, and Hownet synonyms. Those lexicons were used as a simple mapping process. Below, we list the used mapping features.

Type III

- The number of matched positive words
- The number of matched negative words
- The number of matched synonyms
- The number of matched antonyms

3.5 Classification Algorithm

We adopt the SVM (Vapnik, 1995) to learn to classify the testing example. SVM is a kernel-based classifier which can solve non-linear separable problems. Given a set of training examples,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),\ x_i \in \Re^D,\ y_i \in \{+1, -1\}$$

where x_i is a feature vector in D-dimension space of the i-th example, and y_i is the label of x_i either positive or negative. The training of SVMs is to minimize the following objective function (primal form, soft-margin (Vapnik, 1995):

$$\text{minimize}: W(\alpha) = \frac{1}{2}\vec{W}\cdot\vec{W} + C\sum_{i=1}^{n} Loss(\vec{W}\cdot x_i, y_i) \quad (1)$$

The loss function indicates the loss of training error. Usually, the hinge-loss is used (Keerthi and DeCoste, 2005). The factor C in (1) is a parameter that allows one to trade off training error and margin size. To classify a given testing example X, the decision rule takes the following form:

$$y(X) = sign((\sum_{x_i \in SVs} \alpha_i y_i K(X, x_i)) + b) \quad (2)$$

The α_i is the weight of non-zero weight training example x_i (i.e., α_i>0), and b denotes as a bias threshold of this decision. *SVs* means the support vectors and obviously has the non-zero weights of α_i. $K(X, x_i) = \phi(X)\cdot\phi(x_i)$ is a pre-defined kernel function that might transform the original feature space from $\Re^D$ to $\Re^{D'}$ (usually D<<D').

4. Evaluations and results

4.1 Settings

For the classification algorithm, in this paper we adopt the LibSVM (Chang and Lin, 2011) and SVMlight (Joachims, 1998) for training and testing. LibSVM and SVMlight have different strategies for solving multiclass problem. The default setting of LIBSVM is one-versus-one multiclass SVM, while we implement our one-versus-all strategy for SVMlight.

The kernels used in this paper are: 1) polynomial kernel with degree 2 and 2) RBF kernel with Gaussian is 0.03. As seen in (1), the parameter C controls the trained margin and training errors. According to our observations, we set C=1~10 for all experiments. Due to the time constraint, we only submit the result with C=10.

4.2 Results

To validate the parameters, we perform 10-fold-cross validation on the develop data. The develop data comes up with 421 sample text pairs. Table 1 lists the official results of our method in the traditional Chinese MC task. This table only lists the optimal settings: RBF kernel + type I features, polynomial kernel + type I features, and the ensemble method combines typeI and typeII features with RBF and polynomial kernels.

5. Conclusion

Recognizing Inference in Text is an important and new research topic in recent years. Fewer research papers addressed on the Chinese language. This paper presents a hybrid lexical and statistical information-based machine learning framework for RITE task this year. Using only Chinese word segmentation and POS tagging information, this method did not yield the competitive result in official competition result (our 33% v.s. best system report 50%). We also provide an online demonstration of the proposed method[1].

In the future, we plan to integrate more unlabeled data to improve the result. Also, if the parser is available, we will adopt the parse trees.

Table 1: Official results on the traditional Chinese MC task

MacroF1	B-F1	B-Prec.	B-Rec.
32.51	59.21	58.82	59.6
	F-F1	**F-Prec.**	**F-Rec.**
	70.07	61.33	81.71
	I-F1	**I-Prec.**	**I-Rec.**
	8.29	20.27	5.21
	C-F1	**C-Prec.**	**C-Rec.**
	25	20.69	31.58

Acknowledgement:

The authors acknowledge support under NSC Grants NSC 100-2218-E-130-001-

Reference:

[1] I. Androutsopoulos and P. Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research, 38: 135-187.

[2] C. C. Chang and C. J. Lin. 2011. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(27):1-27.

[3] D. Giampiccolo, B. Magnini, I. Dagan and B. Dolan. 2007. The third PASCAL recognition textual entailment challenge. In ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 1-9.

[4] W. C. Huang, S. H. Wu, L. P. Chen, and C. K. 2011. Chinese textual entailment analysis. In Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing.

[5] T. Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In Proceedings of the European Conference on Machine Learning, pages 137-142.

[6] S. Keerthi and D. DeCoste. 2005. A modified finite Newton method for fast solution of large scale linear SVMs. Journal of Machine Learning Research. 6: 341-361.

[7] B. Li, J. Irwin, E. V. Garcia, and A. Ram. 2007. Machine learning based semantic inference: experiments and observations at RTE-3. In ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 159-164.

[8] P. Malakasiotis. 2009. Paraphrase recognition using machine learning to combine similarity measures. In Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, pp. 27–35.

[9] P. Malakasiotis and I. Androutsopoulos. 2007. Learning textual entailment using SVMs and string similarity measures. In Proceedings of ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 42-47.

[10] C. D. Manning and H. Schutze 1999. Fundations of statistical natural language processing. The MIT Press, London.

[11] H. J. Oh, S. H. Myaeng, and M. G. Jang. 2007. Semantic passage segmentation based on sentence topics for question answering. Information Sciences, 177(18): 3696-3717.

[12] M. Tatu and D. Moldovan. 2007. COGEX at RTE3. In ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 22-27.

[13] V. N. Vapnik. 1995. The Nature of Statistical Learning Theory. Springer.

[14] E. M. Voorhees. 2001. Overview of the TREC 2001 question answering track. In Proceedings of the 10th Text Retrieval Conference, 42-52.

[15] S. Wan, M. Dras, R. Dale, and C. Paris. 2006. Using dependency-based features to take the “parafarce” out of paraphrase. In Proceedings of the Australasian Language Technology Workshop, pp. 131-138.

[16] R. Wang and G. Neumann. 2007. Recognizing textual entailment using sentence similarity based on dependency tree skeletons. In ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 36-41.

[17] Yotaro Watanabe and Yusuke Miyao and Junta Mizuno and Tomohide Shibata and Hiroshi Kanayama and C.-W. Lee and C.-J. Lin and Shuming Shi and Teruko Mitamura and Noriko Kando and Hideki Shima and Kohichi Takeda, Overview of the Recognizing Inference in Text (RITE-2) at the NTCIR-10 Workshop, In Proceedings of NTCIR-10 Workshop Meeting, 2013.

[18] Y. C. Wu, Y. S. Lee, and J. C. Yang. 2008. Robust and efficient multiclass SVM models for phrase

[1] http://120.96.128.186/ritc_ct

pattern recognition. Pattern Recognition, 41(9): 2874-2889.

[19] Y. C. Wu, J. C. Yang, Y. S. Lee, and S. J. Yen. 2010a A Sparse L2-Regularized Support Vector Machines for Large-scale Natural Language Learning. Proceedings of 6th Asia Information Retrieval Symposium (AIRS), pp. 340-349.

[20] Y. C. Wu, J. C. Yang, Y. S. Lee, and S. J. Yen. 2010b. An Integrated Deterministic and Nondeterministic Inference Algorithm for Sequential Labeling. Proceedings of 6th Asia Information Retrieval Symposium (AIRS), pp. 221-230.

[21] Y. C. Wu, J. C. Yang, Y. S. Lee, and S. J. Yen. 2010c. Chinese Word Segmentation with Conditional Support Vector Inspired Markov Models. Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP'2010), pp. 228-233.

[22] Y. C. Wu, C. J. Lee, and Y. C. Chen, "MCU at NTCIR: A Resources Limited Chinese Textual Entailment Recognition System," In Proceedings of the 9th NTCIR workshop meeting on evaluation of information access technologies, 2011.