MCU at NTCIR: A Resources Limited Chinese Textual Entailment Recognition System

Yu-Chieh Wu
Dept. of Communication and Management
Ming-Chuan University
Taipei, Taiwan
wuyc@mail.mcu.edu.tw

Chung-Jung Lee
Financial Department
Ming-Chuan University
Taipei, Taiwan
cjlee@mail.mcu.edu.tw

Yaw-Chu Chen
Dept. of Communication and Management
Ming-Chuan University
Taipei, Taiwan
ycchen@mail.mcu.edu.tw

Abstract

Recognizing inference in text is the task of finding the textual entailment relation 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. In this paper, we go different line. We propose a pattern-based and support vector machine-based trainable text entailment tagging framework under the condition of part-of-speech tagging information is available. 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. In terms of accuracy, our method achieves 53.6% for Traditional Chinese MC task (second place) and 55.4% for Traditional Chinese BC task. After the correction, our method in BC task is 67.9% with the same setting.

1. Introduction

Text mining 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 has been to create Chinese/US/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 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 propose a machine learning-based textual entailment recognition framework with very limited resources. Only Chinese word segmentation and POS tagging are required. 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. To further enhance the result, we compare several kernels and settings. As our experiments, we found the polynomial kernel SVM achieve the optimal solution for our need. In short, we achieve the second place in the traditional Chinese MC task.

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. Task Analysis

As in paraphrasing, textual entailment of RITE task needs to develop a system which can identify five different relationships between given pair of text fragments. The first type is the forward entailment where the hypothesis includes the means of the text, i.e. $h \rightarrow t$ but t cannot infer h, reversely. Here lists an example pair of forward type.

Forward

h: 1970 年畢蘭德拉國王和皇后艾斯瓦利亞結婚
t: 畢蘭德拉和艾斯瓦利亞是夫妻

Alternatively, the reverse entailment means the t can infer h, but h cannot infer t. In other words, t completely includes the entire meanings h. An example lists below.

Reverse

h: 丁磊是網易的創辦人
t: 丁磊 1997 年 6 月創立網易公司

For the third type, both t and h can be inferred each other. That means, $h \rightarrow t \wedge t \rightarrow h$. Below is an example of the bidirection entailment text fragments pair.

Bidirection

h: 二次世界大戰時日本廣島遭投原子彈
t: 廣島在二次世界大戰時遭原子彈轟炸

Both independence and contradiction indicate that there is no any single textual entailment relation between the text pair. The independence means the given text pair express two different meanings, while the contradiction shows that the fact of h is true in h but false in t. Below list the two entailment types.

Independence

h: 籐森 2000 年遭秘魯國會免職
t: 籐森曾任秘魯總統

Contradiction

h: 4 月 26 日當天，有 3 種版本的 CIH 同時發病
t: CIH 病毒，4 月 26 日只有 1.2 和 1.3 版發作

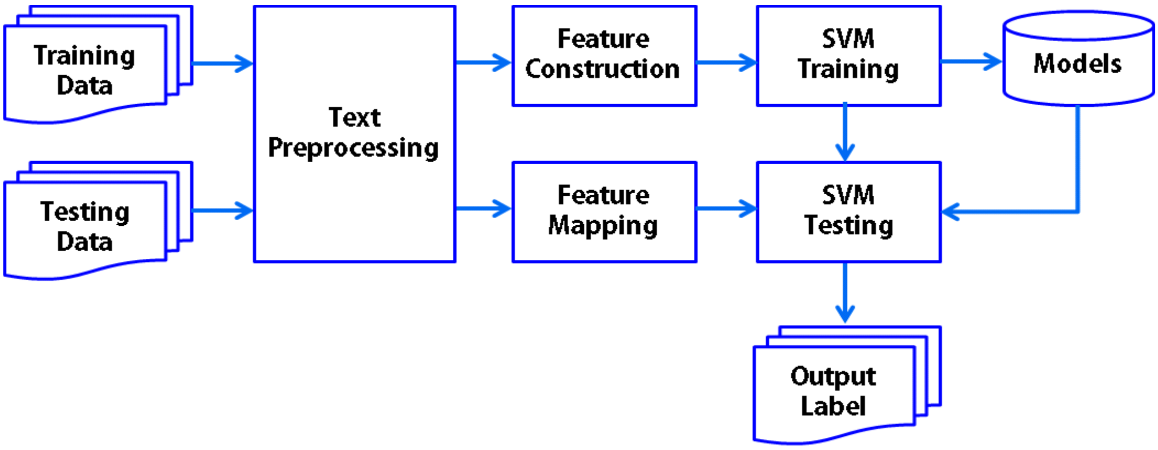

Figure 1: Overall System Flowchart

As shown in the contradiction type, h indicates that there are 3 CIH virus types. However, t merely reveals that there are only two CIH virus types. Obviously, *h* and *t* are contradiction with each other.

3. Methodology

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 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.

3.1 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., 2008, 2010a, 2010b, 2010c).

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 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 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. The ensemble method combines four methods with weighted voting. The weight is mainly determined by the accuracy on the validation data.

Table 1: Official results on the traditional Chinese MC task

Method	Validation data	Testing Data
Official method	56.11%	53.60%
RBF kernel	53.44%	51.44%
Polynomial kernel	56.11%	53.60%
Ensemble method	56.25%	53.80%

In the BC track, we originally convert {*F*, *R*, *B*, *C*} as true entailment label, while *I* is treated as false. Owing to this mistake, we only achieve 55.40% accuracy in the BC task. By converting {*F*, *B*} to true (has entailment relation), and {*R*, *C*, *I*} to false, we obtain the correct result. Table 2 lists the correct result on the traditional Chinese BC task. We also compare the two SVM tools from the aspect of view of multiclass problem. Clearly, SVMlight (one-versus-all) is weak in MC task in comparison to LIBSVM (one-versus-one). However, SVMlight is very suitable for BC task which achieves 67.89% accuracy in binary problem. In the MC task, the LIBSVM has much better prediction power than SVMlight where the accuracy is 53.60%.

Table 2: Performance comparison between SVM-light and LIBSVM

Method	MC	BC
SVMlight	51.33%	67.89%
LIBSVM	53.60%	55.40% (official) 66.88% (correct)

Third, we run the experiments on the testing data using different features. Table 3 shows the result on the MC task. Obviously, the polynomial kernel with type I feature obtains the optimal solution. Type I feature significantly outperforms the type II.

Table 3: The effect of different feature type

Method	Accuracy on testing data
Type I + RBF	51.44%
Type II + RBF	38.66%
Type I + Polynomial	53.60%
Type II+ Polynomial	23.44%

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 achieves the second place in official competition result. As summary, it achieves 53.60% in MC task. After classifier fusing, this method can be even enhanced to 53.80%.

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.

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] Y. C. Wu, Y. S. Lee, and J. C. Yang. 2008. Robust and efficient multiclass SVM models for phrase pattern recognition. Pattern Recognition, 41(9): 2874-2889.

[18] 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.

[19] 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.

[20] 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.