NWNU Minimum Information Recognizing Entailment System for NTCIR-11 RITE-3 Task

Zhichang Zhang, Dongren Yao, Longlong Mao, Songyi Chen

School of Computer Science and Engineering, Northwest Normal University, Lanzhou, China
zzc@nwnu.edu.cn;wade330628704@163.com;
775919087@qq.com;snail200x@163.com

Abstract

This paper describes our work in NTCIR-11 on RITE-3 Binary-class (BC) subtask and Multi-class (MC) subtask in Simplified Chinese. We proposed a textual entailment system using a hybrid approach that integrates many features. The performance of the proposed method in the formal run achieved Macro-F1's of 59.71% in BC subtask and only 23.19% in MC subtask

Key words

Textual entailment, Machine Learning, RITE.

1 Introduction

The Recognizing Inference in TExt (RITE) challenge focuses on detecting the directional entailment relationship between pairs of text expressions, denoted by T (the entailing "Text") and H (the entailed "Hypothesis"). We say that T entails H if human reading T would typically infer that H is most likely true [1].

NWNU team participated in NTCIR-11 RITE-3 Binary-class (BC) subtask and Multi-class (MC) in Simplified Chinese (CS). We submitted 5 official runs for BC subtask and 3 runs for MC subtask. Our system focuses on the minimum information between two texts, and we find out those information help our system recognize the entailment relationship more effectively.

The rest of the paper is arranged as follows: section 2 describes the features and algorithms employed in our system in detail; section 3 presents and discusses the official evaluation; section 4 concludes this paper with a description of the future work.

2 System description

There will be four main modules in our system, i.e. preprocessing, feature extraction, classifier and amending module. Figure 1 show more details in our system:

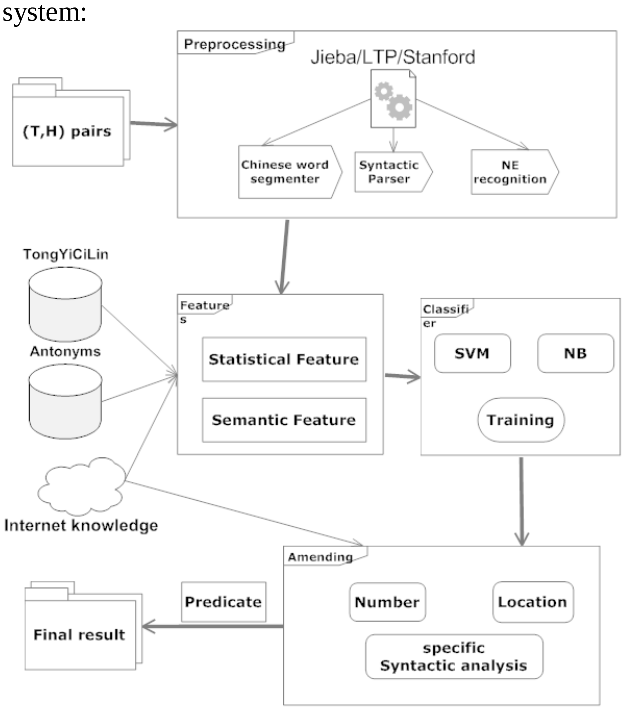

Figure1. System architecture

2.1 Preprocessing module

This module uses hybrid NLP resources and tools for basic processing like word segment, POS, and NE recognition. We use Jieba as our word segment tool, and LTP[1] tool for POS tagging, dependency syntactic parsing and Stanford classifier for recognizing on named entity.

2.2 Feature extraction

Because of the complexity on textual entailment, many factories should be taken into consideration. Some traditional features including statistical and lexical semantic features are used in the statistical machine learning models to classify whether *T* in a given text-hypothesis pair entail *H*. The following table

---

[1] http://www.ltp-cloud.com/demo/

1 and table 2 illustrate these features and equations for them.

Table 1. Statistical features

Feature Name	Comment	Formula
Word overlap	The overlap of word between two texts	$E_1=\|T\wedge H\|/\|H\|$ $E_2=\|T\wedge H\|/\|T\|$ $E=(2*E_1*E_2)/(E_1+E_2)$
Length difference	Using text length to distinguish entailment direction	$Lt(T,H)=\|Len(T)-Len(H)\|$
Cosine similarity	Representing the text pair as vectors, then calculating their cosine similarity	$Sim_{cos}(T,H)=\frac{\sum_{i=1}^{n} t_i * h_i}{\sqrt{\sum_{i=1}^{n} t_i} * \sqrt{\sum_{i=1}^{n} h_i}}$ n is vector dimensions

Table 2. Lexical semantic features

Feature Name	Comment	Formula
Tongyicilin semantic similarity [2]	Using Tongyicilin to calculate the similarity between different words	See Equation 1
The number of antonyms	Using the Web resource to count the number of antonyms	None
The number of negative words	Combining the number of antonyms to assist the decision	None
The overlap of named entity	Named entities can show the text topics in a way	$T_{NE}=\|T\wedge H\|/\|H\|$ $H_{NE}=\|T\wedge H\|/\|T\|$ $L_{NE}=(2*T_{NE}*H_{NE})/(T_{NE}+H_{NE})$

$$Sim=\frac{1}{2}\left[\frac{\sum_{i=1}^{m} max\{sim_w(w_{1i},w_{2j})\|1\le j\le n\}}{m}+\frac{\sum_{j=1}^{n} max\{sim_w(w_{1i},w_{2j})\|1\le i\le m\}}{n}\right] \quad (1)$$

These two kinds of features can help our system solve some problems as：

<pair id="98">

<t1>1999 年，美国国会图书馆将罗马假期评为“文化杰出奖”，并收入“国家级收藏”。</t1>

<t2>1999 年，罗马假期被美国国会图书馆评为“文化杰出奖”，并收入“国家级收藏”。</t2>

</pair>

With a high level character similarity, our system will treat them as ‘entailment’, and the fact is same as the judgment. But these features can’t catch the syntactic structure. So the traditional feature will misclassify many pairs, and we’ll talk about it in amending module.

2.3 Classifier

Different machine learning methods produce different results. SVM which is regarded as the best supervised learning method in general text classification didn’t obtain the best result during the training phase, but got the best score on testing set. The Naïve Bayes based on Gaussian distribution had the best performance on training set, but didn’t do as good as SVM during the testing phases. We also believe in BC subtask, the voting format won’t help system do better in textual entailment, Maybe it works in MC subtask. So, we use single machine learning method to predicate the result, and the open source tool scikit-learn [4] is employed for classification in this system

2.4 Amending module

As we described in feature extraction, traditional features catch some factories in textual entailment. But the level of syntactic structure wouldn’t be analyzed with these features. For example:

<pair id="241">

<t1>日本于 2005 年发行上映的动画电影《蒸气男孩》，其故事背景以英国 1851 年万国博览会为主。</t1>

<t2>英国于 2005 年发行上映的动画电影《蒸气男孩》，其故事背景以日本 1851 年万国博览会为主。</t2>

</pair>

<pair id="594">

<t1>南北朝陶弘景所著的《名医别录》概括的琥珀三大功效为：去惊定神，活血散淤，利尿通淋。</t1>

<t2>南北朝陶弘景所著的《名医别录》概括的琥珀三大功效为：去惊定神，利尿通淋，活血散淤。</t2>

</pair>

With one hundred percent character similarity, without amending module, machine learning algorithm predicates both of them ‘entailment’. In fact, in the id=”241”, *t*1 don’t entail *t*2, *t*1 and *t*2 are contradicted. So we need to analyze the syntactic structure on each text.

Furthermore, numbers occur in sentences in different ways (Arabic numerals, Chinese numerals, or numbers combing with Chinese character). We need to effectively find out whether the different formats affect the relationship between two texts.

The recognized Named Entity such as locations and institutions should be taken into consideration separately. There are too many relationships with two named entities like synonymy, hyponym and independence. Relationships between two NE are significant to the determination on whether or not two pairs can entail.

3 Official Evaluation

We submitted five results of BC and three results of MC to NTCIR-10. Our system only focused on the BC subtask, and the relationship in MC we didn’t take it into consideration.

3.1 Formal run results

The formal run results on BC of our system are shown in Table 3 and Table 4. Each team can run many times, we only use their best Macro-F value as their results.

Table 3. Result of the formal run on BC[5]

Participants	Macro-F	ACC	Y-F	N-F
BUPT	61.51	62.33	67.15	55.86
NWNU	59.71	59.75	60.95	58.47
III&CYUT	56.75	56.75	57.07	56.42
WHUTE	53.48	54.58	60.65	51.49
Yamraj	49.24	49.25	48.69	49.79
ASNLP	44.95	51.50	63.94	25.95
IMTKU	42.80	53.25	67.25	18.34
JAVN	42.32	51.17	64.91	19.73
WUST	39.14	52.25	67.39	10.89

According to the table 3 and table 4, we found that Y-F value was much lower than B* team because of Y-Rec value. That means words similarity based on Internet added in round four and five confused our system to recognize the entailed pairs. But this feature solved some problems like this:

<pair id="962">

<t1>爱因斯坦出生后不久，便于 1880 年举家迁往慕尼黑。</t1>

<t2>爱因斯坦出生后不久，便于 1880 年举家搬到慕尼黑。</t2>

</pair>

When we use the words similarity based on Internet, ‘迁往’and ‘搬至’will treat as synonyms. This method would be a good supplement for Tongyicilin.

Table 4. Result of the formal run on BC about NWNU

Participants	Macro-F	ACC	Y-F	Y-Prec	Y-Rec	N-F	N-Prec	N-Rec
B*-CS-SVBC-05	**61.51**	62.33	67.15	59.54	77.00	55.86	67.45	47.67
B*-CS-SVBC-04	61.42	62.83	68.81	59.28	**82.00**	54.02	70.81	43.67
B*-CS-SVBC-02	60.82	62.33	68.52	58.85	82.00	53.11	70.33	42.67
B*-CS-SVBC-01	60.54	62.08	68.34	58.66	81.83	52.75	69.97	42.33
B*-CS-SVBC-03	60.54	62.08	68.34	58.66	81.83	52.75	69.97	42.33
N*-CS-SVBC-05	59.71	59.75	60.95	59.18	**62.83**	58.47	60.39	56.67
N*-CS-SVBC-04	58.83	58.83	58.90	58.80	59.00	58.76	58.86	**58.67**
N*-CS-SVBC-03	58.03	59.00	64.40	56.91	74.17	51.67	62.92	43.83
N*-CS-SVBC-02	51.83	55.00	64.19	53.30	80.67	39.46	60.27	29.33
N*-CS-SVBC-01	45.82	51.75	63.74	51.05	84.83	27.90	55.17	18.67

We submitted five results in five different ways. The first result with traditional methods achieved a good performance during in RITE2 testing set. But when we used this system to predict relationship with training set, the result showed pretty ugly. We found new training set pay more attention to the relationship between lexical, e.g. abbreviation, antonym, hypernymy and so on. Traditional features can’t represent the relationship between lexical, kinds of dictionaries and knowledge should be added into entailment system. The second and third results with different machine learning methods increase our system via using amending module. And for the last two results, we use some Knowledge through baidu.

For MC subtask, we didn't catch the nature of this problem. The method only based on BC subtask has be proved that it is not effective.

3.2 Discussion

There is much room left to further improve the NWNU system. Several directions for further improvement are list as below:

1. With more fine sorted relationship between two texts, the problem will refine into different situation. One typical category be solved, our system will achieve a better performance.
2. World knowledge from Wikipedia or Baike is important to a whole entailment system. How to establish an effective hierarchy to combine all knowledge and to use these knowledge duly should be a vital factor to our system.
3. Our system would pay more attention to MC subtask, more features would be added into
4. More concentration on syntactic analysis should be done in our system.

4 Conclusion

This paper presents the NWNU system in the NTCIR-11 RITE-3 challenge. We participate in the BC, MC task on Simplified Chinese. Different linguistic level features and machine learning methods are used in our system to improve the recognition accuracy.

Our future work is two-fold. The judgment on entailment pairs is not successful, therefore we should find new features or new methods to express entailment relationship better. Meanwhile we also need to focus on the multi-direction in Chinese textual entailment recognizing, and this challenge requires entailment system developed in a more robust and reasonable way.

5 Acknowledgments

This work was sponsored by the National Natural Science Foundation of China (No. 61163039, No. 61163036, No. 61363058) and the Young Teacher Research Ability Enhancement Project of Northwest Normal University of China (NWNU-LKQN-10-2). We thank the anonymous reviewers for their insightful comments.

6 References

1. Dagan I. and Glickman O. Probabilistic textual entailment: generic applied modeling of language variability[C]//In PASCAL Workshop on Learning Methods for Text Understanding and Mining, Grenoble, France. (2004).
2. Tian jiu-le, Zhao Wei, Words Similarity Algorithm Based on Tongyici CiLin in Semantic Web Adaptive Learning System [J].Journal of Jilin University，2010,28(6),602-608
3. Watanabe Y, Miyao Y, Mizuno J, et al. Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10, Proceedings of the 10th NTCIR Conference. 2013.
4. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011
5. Suguru Matsuyoshi and Yusuke Miyao and Tomohide Shibata and Chuan-Jie Lin and Cheng-Wei Shih and Yotaro Watanabe and Teruko Mitamura, Overview of the NTCIR-11 Recognizing Inference in TExt and Validation (RITE-VAL) Task, Proceedings of the 11th NTCIR Conference, 2014