FudanNLP at RITE 2011: a Shallow Semantic Approach to Textual Entailment

Ling Cao, Xipeng Qiu and Xuanjing Huang

11210240044, xpqiu, xjhuang@fudan.edu.cn

Abstract

RITE is a task recognizing logic relations between texts. This paper presents FDCS's approach on NTCIR9-RITE Chinese simplified BC & MC subtasks. Our system is built on a machine learning architecture with features selected on shallow semantic methods, including named entity recognition, date & time expression extraction, words overlap and negation concept recognition. FudanNLP is wildly used by our system on NLP procession and feature extraction. System gets accuracy as 76% on BC subtask and 58.5% on MC subtask, separately.

1. Introduction

The field of recognizing textual entailment has generated growing interest in the past few years [1]. As a common phenomenon of natural language, same meanings can always be expressed by different words. The example below shows the equivalent of "登陆" and "引进" (both means "landed"), "内地" and "大陆" (both means "the mainland of China"):

Text1: 1998年登陆内地的《泰坦尼克号》，曾获得3.2亿元票房。(《Titanic》, landed to the mainland of China in 1998, got a box office as 320 million RMB.)
Text2: 1998 年引进大陆的《泰坦尼克号》，其票房是 3.2 亿元。(The box office of《Titanic》, which landed to the mainland of China in 1998, is 320 million RMB.)

For the language viability problem, recognizing textual entailment is essential to many natural language processing applications, such as Question Answering (QA), Information Extraction (IE), machine translation (MT) evaluation, etc.

RITE [2] is a task to recognize semantic relation between texts, that is, given two text fragments, recognizes whether one can be entailed by the other. And the multi-classification subtask is to find more detailed relations between two text fragments: Forward entailment, reverse entailment, bidirectional entailment (same meaning), independence and contradiction.

The system of the FDCS Group which has taken part in the 2011 RITE is a proposal towards the resolution of the language viability problem. By using FudanNLP[1], our current approaches explore a method by finding common grounds between pairs with same semantic relation. With machine learning methods, system extract features from training sets on shallow semantic level.

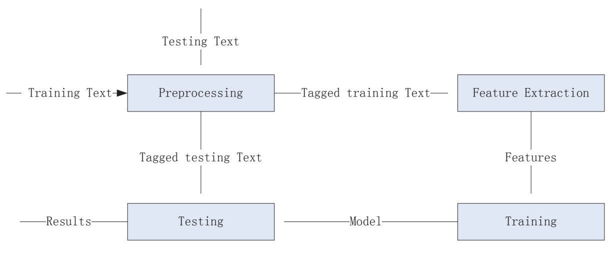

Fig. 1 System architecture

The rest of the paper is organized as follows: Section 2 presents a brief introduction to our system architecture. Then we describe feature selection in details in section 3. Finally, we

[1] http://code.google.com/p/fudannlp/

present our experiments in section 4 and conclude in section 5.

2. System Overview

The framework of our system is presented in Fig. 1. In our system, training set is sent to preprocessing module initially.

The function of preprocessing module is as follows:

(1) Text extraction.
(2) Tokenizing, POS tagging and parsing.
(3) Named entities recognition.
(4) Date and time recognition.
(5) Number recognition.
(6) Dependency analysis.

Starting from text pairs in the RITE XML format, our system uses RITE-SDK[2] to extract texts from training set. After text extraction, FudanNLP is used to preprocess text. And then, annotated training text is sent to feature extraction module. Feature selection of our current approach will be introduced in next section. After feature extracting, system uses LibSVM [3] to get the training model from the feature vectors of training set. After training model is generated, system extract feature vectors from test set in the same way as training set. In the last step, training model classifies text pairs into different categories, according to the feature vectors of them.

3. Feature Selection

With the annotated data obtained in previous step, system computes the value of feature vectors. The following features are constructed for use in the training of the model, and producing entailment predictions.

Named Entity Cover Based on the guideline of judging whether t1 entails t2 or not, t2 that introduce entities which are not mentioned by t1 indicates t2 cannot be entailed by t1, and vice versa. For example:

Text1: 1997年香港回归中国。(Hong Kong was given back to China in 1997.)
Text2: 香港的主权和领土是在1997年由英国归还给中国的。(Hong Kong was given back to China by Britain in 1997.)

The named entity "英国 (Britain)" introduced in t2 is never mentioned in t1, so t1 cannot entail t2. But all the named entities introduced in t1, including "香港 (Hong Kong)" and "中国 (China)", are mentioned in t2, it is likely that t1 can be entailed by t2. In fact, the semantic relation of the two pairs is "R" (reverse entailment).Therefore, we select the named entity coverage situation of text pairs as a feature. We use FudanNLP named entity recognition module to extract entities and compare whether each entity is mentioned in both two text fragments.

Date, Time & Number Cover Just the same as named entities, if t1 and t2 happened in different periods of time, they can hardly entailed by each other. And if we know when t1 happened while the time t2 happened is not mentioned, then t2 cannot entail t1. The example below may better explain the facts:

Text1: 普京于2000年3月当选为俄罗斯总统,并于2004年3月14日在大选中再度竞选获胜，连任成功。(Putin was elected president of Russia in 2000 March, and won the election again in March 14, 2004.)
Text2: 俄罗斯前总统叶利钦1999年12月31日宣布辞去总统职务,并宣布由普京代行总统职务。(Former Russian President Yeltsin announced his resignation from the presidency in December 31, 1999. And assist Putin as president.)

Without deep semantic understanding, the date & time information indicates that t1 and t2 are telling things happened in different periods of time, so they tend to be independent. And of course, if t1 and t2 are almost telling the same

[2] http://code.google.com/p/rite-sdk/

thing happened in different periods of time, they can be contradictory.

Number information of text pairs is also worth paying attention to. For example:

Text1: 2001年世界游泳赛，郭晶晶在女子三米板跳水项目上以三四七点三一封后(In 2001 World Swimming Championships, Guo Jingjing wins as a score of three hundreds forty seven point three.)
Text2: 郭晶晶以 381.75 分的总成绩夺得2001年世界游泳锦标赛女子三米跳板金牌(In 2001 World Swimming Championships, Guo Jingjing wins as a score of 381.75.)

The score Jingjing Guo get in t1 and t2 is not identical, so they are contradictory.

What worth to say is that a same date/time/number can be expressed in variable type. For instance, the following expressions are all means time 1995-8-18 21:30:

8/18/1995 21:30 1995/8/18 21:30 1995.8.18 9:30 PM 一九九五年八月十八日二十一点三十分 95 年 8 月 18 号晚上 9 点半

By using FudanNLP time expression recognition module, all above can be recognized as 1995-8-18 21:30. And also, this module is able to translate Chinese number into Arabia digit, ie, “三千零七十五点三一” into “3075.31”.

Word Length & Word Overlap This simplest feature may also worth selection. If t1 is much longer than t2, it seems that t1 gives more information than t2, so t2 cannot entail t1. And if t1 and t2 are not similar enough, they may talk about different things.

Negation & Antonym The existence of negation words and antonyms may indicate contradictory between text pairs. So system recognizes whether one of the texts in a text pair contains negation concept, or both of them do, as a feature.

System uses Hownet [4] antonym dictionary to find antonym words and construct a list of rules to find negation words. for example:

Text1: 网络泡沫化指的是网络产业因过度膨胀而形成泡沫。(Network bubble refers to the enlarge of network industry.)
Text2: 网络产业因过度紧缩而泡沫化(Network bubble refers to the shrink of network industry.)

Hownet explain “膨胀” as “{enlarge|扩大}” and “紧缩” as “{shrink|缩小}”. And by looking up Hownet antonym dictionary, system recognize “膨胀” and “紧缩” as antonyms . So with Hownet, system can recognize antonyms on semantic level.

4. Experiments and Results

We evaluates our method on the dataset of RITE 2011. The only difference between two runs is that run2 extract negation & antonym feature while run1 do not.

Table 1. Overall RITE run result

	Run 1	Run 2
BC subtask	0.746	0.76
MC subtask	0.58	0.585

Table 2. Accuracy rate of each tag in BC subtask

	Run 1	Run 2
Y	0.871	0.848
N	0.521	0.604

Table 3. Accuracy rate of each tag in MC subtask

	Run 1	Run 2
F	0.624	0.634
R	0.824	0.747
B	0.873	0.831
C	0.189	0.351
I	0.314	0.3

Table 1 shows the overall results of our two run in two subtasks. Table 2 and Table 3 show more detailed accuracy rate of each tag in BC & MC subtasks, respectively. Comparing run1 and run2 in MC subtask, it turns out that by extracting negation & antonym feature, run2 achieves much better results than run1 in the classifying of contradictory relation, as the accuracy of tag “C” doubled from run1 to run2. But the overall accuracy does not improved obviously as the accuracy of tag “R”, “B” and “I” are all declined. By analyzing the text set, we find out that although the extraction of negation & antonym feature improves system ability of recognizing negation concepts, it also mislead the system into wrong way, for example:

Text1: 矿物质类固醇多在医师处理休克等一些紧急状况使用。(Mineral steroid in physician treatment of shock and other emergency use.)
Text2: 类固醇在药理学上分为皮质类固醇及矿物质类固醇,后者在医生处理休克等一些紧急状况才会使用,一般民众极少有机会接触(Steroids in pharmacology as corticosteroids and minerals of steroids, the latter in the doctor treating shock and other emergency will be used, people rarely have the opportunity to contact)

System finds out negation concept of the text pair above, as “多” in text1 and “少” in text2. But in fact, t1 and t2 do not contains negation concepts at all, as “多” and “少” are modified different things. So without syntactic analysis and deep semantic analysis, system may hardly solve this kind of problem.

5. Conclusions

This paper reports FDCS system on RITE. The system is built on machine learning framework with features selected on shallow semantic methods. FudanNLP is wildly used by our system on NLP procession and feature extraction. System gets an accuracy as 76% on BC task and 58.5% on MC task. During experiments, it turns out that many problems cannot be solved without deep understanding of texts, so we will focus on deep semantic understanding of textual entailment in our future research.

Acknowledgement

This work was (partially) funded by NSFC (No. 61003091 and No. 61073069) and Shanghai Committee of Science and Technology (No. 10511500703).

References

[1] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan,Bill Dolan. 2007. The Third PASCAL Recognizing Textual Entailment Challenge, in Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, Czech Republic.

[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] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[4] Zhendong Dong, Qiang Dong. Hownet. http://www.keenage.com/.