Recognizing Textual Entailment Using Multiple Features and Filters

Yongmei Tan Minda Wang Xiaohui Wang Xiaojie Wang

School of Computer Science
Beijing University of Posts and Telecommunications
Beijing 100876, China
{ymtan, mindawang, xiaohuiwang, xjwang}@bupt.edu.cn

Abstract

Textual entailment among sentences is an important part of applied semantic inference. In this paper we propose a novel technique to address the recognizing textual entailment challenge, which based on the distribution hypothesis that words that tend to occur in the same contexts tend to have similar meanings. Using the IDF of the overlapping words between the two propositions, we calculate the similarity between the two given propositions to infer the likelihood of entailment and then filter the results inferred. We evaluate our model on NTCIR-11 RITE dataset and then show how a combination of multiple features and filters can significantly improve the performance of recognizing textual entailmentover the best performers in those years. Our approach advances state-of-the-art Simplified Chinese NTCIR-11 RITE.

1. Introduction

Recognizing Textual Entailment (RTE) (Dagan et al., 2006) is a task to detect whether one Text (T1) can be inferred (or entailed) by another Text (T2). Being a challenging task, it has been shown that it is helpful to applications like question answering (Harabagiu and Hickl, 2006), summarization (Barzilay et al., 1999) and information retrieval (Anick and Tipirneni, 1999). RTE (Bentivogli, et al., 2011), a series of evaluations on the developments of English Textual Entailment (TE) recognition technologies, have been held seven times up to 2011. In the meanwhile, TE recognition technologies in other languages are also underway (Shima, et al., 2013, Huang et al., 2013).

The main hypothesis in this work is that Harris' Distributional Hypothesis, which states that words that occurred in the same contexts tend to be similar (Harris, 1985).

We model RTE task as the following 4 steps. Firstly, we preprocess the data, including temporal expressions normalization, numerical expressions normalization and character expressions normalization. Secondly, T1 and T2 are processed by segment, POS tagging, named entity recognition and co-reference resolution. Thirdly, using external resource, we extract multiple features related to T1 and T2 to build a classifier. Finally, we calculate the entailment score between T1 and T2 and then predict the entailment relation, either "YES" or "NO"by several filters.

We make the following contributions:

1) Different from traditional approaches, we present a novel framework for textual entailment recognition, which focus on multiple features and filters. (Section 3).
2) We achieved the best results on Simplified Chinese NTCIR-11 RITE BC subtask (Section 5).

2. Related Work

In recent years, many researchers have focused on RTE. They have developed lots of methods based on logical inference (Hickl and Bensley, 2007; Clark and Harrison, 2009), similarity between dependency parse trees (Bar-Haim et al., 2009) or similarity between syntactic graphs (Padó et al., 2009). Such previous works have made significant progress in RTE (Sammons et al., 2010) beyond a smart lexical baseline (Do et al., 2009). The top 3 systems (Tsuchida and Ishikawa, 2011; Yokote et al., 2011; Tan et al., 2011) in RTE7 are primarily basically lexical-level matching approaches.

3. Recognizing Textual Entailment Framework

The framework of our method is shown in Figure 1. The basic components are preprocessing, processing, modeling, and filtering.

3.1 Preprocessing

There are many different word expressionsin T1 and T2, while they express the same meaning. In order to improve the performance of our model, we complete temporal expressions normalization, numerical expressions normalization and character expressions normalization in this step.

3.2 Processing

We use the Stanford CoreNLP[1] to complete word segment, POS tagging, named entity recognition and co-reference resolution.

[1]http://nlp.stanford.edu/software/corenlp.shtml

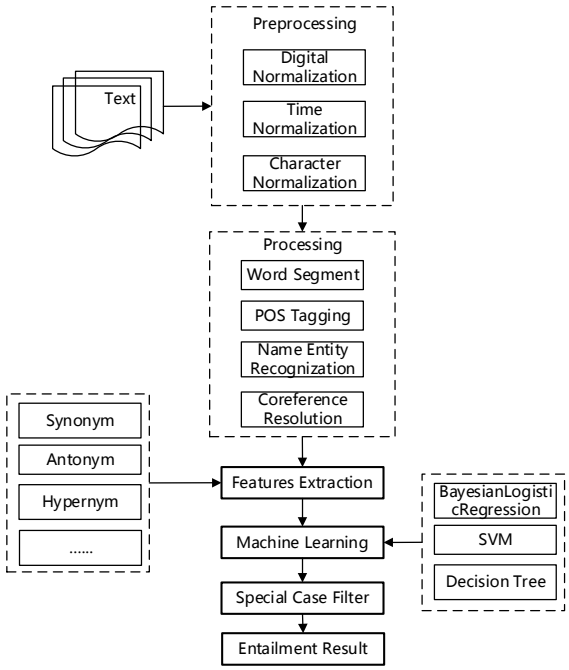

Figure1: Recognizing Textual Entailment Framework

3.3 Modeling

A summary of the features is listed in Table 1.

Table 1: Feature Set for Recognizing Inference in Text

ID	Description
0	The similarity measure *SM(T1,T2)* between T1 and T2
1	True if the number of the negatives in T1 is equal to those of T2
2	Number of the named entities appearing in T1 but not appearing in T2
3	True if the number of the antonyms in T1 is equal to those of T2
4	True if antonyms appearing in T1 and T2
5	True if the word that begins with "<" and ends with ">" appearing in T2, but not appearing in T1
6	The difference between T1 length and T2 length
7	False if NEs of the T1 and T2 are consistent
8	True if the only one word is different between T1 and T2
9	True if the same words between T1 and T2 but the word order is different between T1 and T2
10	True if the temporal expressions appearing in T1 and T2 are inconsistent
11	True if the numerical expression appearing in T2 but not appearing in T1
12	True if "等/ and so on" appearing in T1 and T2?
13	True if the difference is that "支持、希望/support, hope" between T1 and T2
14	True if the difference is that "可能、或许/ can, maybe"
15	True if the difference is that "被认为/ be considered" between T1 and T2
16	True if the difference is that "或/ or" or "和/ and"
17	True if the title / position appearing in T2, but not appearing in T1
18	True if the difference is that number expression between T1 and T2
19	True if the difference is that "第一/first" between T1 and T2
20	True if the difference is that synonyms between T1 and T2
21	True if T1 and T2 are same
22	True if the difference is that location between T1 and T2

The similarity measure $SM(T_1, T_2)$ between T_1 and T_2 is calculated as below:

$$SM(T_1, T_2) = \frac{\sum_{w \in \{T_{1w} \cap T_{2w}\}} (IDF(w))^{\alpha}}{\sum_{w \in T_{2w}} (IDF(w))^{\alpha}}$$

Here T_{1w} and T_{2w} denote the word sets of T_1 and T_2 respectively. α is the exponent (we use 1, 2 and 3). $IDF(w)$ is the inverse document frequency of word w. It is defined as follows.

$$IDF(w) = \begin{cases} \frac{N}{f(w)} & \text{if } w \text{ appears in training data} \\ \beta & \text{if } w \text{ is an unkown word} \end{cases}$$

where N is the total number of sentences in the training data and β is a very small nonzero number obtained from the training data.

Based on the above features, our models are trained by Support Vector Machines (SVM), Logistic Regression, Bayes Network (Bayesnet) and Decision Tree (DT).

3.4 Filtering

The filtering mechanism conservatively modifies the results of T1 and T2 pairs detected by the above models. That is we discard such pairs if the model predicts false-positive pairs caused by the classifier and choose such pairs if the model predicts false-negative pairs caused by the classifier with high confidence.

4. Experiments

4.1 Experimental Settings

NTCIR-11 RITE-VAL is a generic benchmark task that addresses common semantic processing needs in various NLP/Information Access research areas, which includes two subtasks: Fact Validation (Search) and System Validation (Unit test). In System Validation, there are four subtasks, i.e. Binary Class (BC),Multi Class (MC), Entrance Exam and RITE4QA. We just focus on BC subtasks.

The systems were evaluated by macro-F1score which is defined by

$$macroF1 = \frac{1}{C}\sum_{c \in C} F1_c = \frac{1}{|C|}\sum_{c} \frac{2 \times P_c \times R_c}{P_c + R_c}$$

where C is the set of classes and P_c and R_c is a precision value and a recall value for the class c. Precision and recall are defined as follows.

$$P = \frac{N_{correct}}{N_{predicted}}$$

$$R = \frac{N_{correct}}{N_{target}}$$

The statistics of the data are shown in Table 1. There are more positive pairs than negative pairs in training data and the number of test data doubles that of training data.

Table 1: Simple Statistics of the Simplified Chinese NTCIR-11 RITE Data

BC	Y	N	Total
CS (Training data)	370	211	581
CS (Test data)			1200

4.2Experimental Results

We submitted the following runs. The differences between feature set 1 (FS1) and feature set 2 (FS2) is that we update feature 9 (Table 1) in FS2.

The results are shown in Table 2.

BT1: FS1 + Libsvm + Filtering
BT2: FS2 + Libsvm + Filtering
BT3: FS2 + Logistic + Filtering
BT4: FS2 + Bayesnet + Filtering
BT5: FS2 + DT + Filtering

Table 2: Results of Simplified Chinese NTCIR-11 RITE

Run	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
BT1	60.54	62.08	68.34	58.66	81.83	52.75	69.97	42.33
BT2	60.82	62.33	68.52	58.85	82.00	53.11	70.33	42.67
BT3	60.54	62.08	68.34	58.66	81.83	52.75	69.97	42.33
BT4	61.42	62.83	68.81	59.28	82.00	54.02	70.81	43.67
BT5	**61.51**	62.33	67.15	59.54	77.00	55.86	67.45	47.67

The experiment performs well on the dataset, achieving a MacroF1 of 61.51 from the above table.

5. Conclusion

This paper proposes a novel technique focused on recognizing textual entailment challenge. Experiments on Simplified Chinese NTCIR-11 RITE show the effectiveness of using statistical model in conjunction with filters: 61.51% MacroF1 is achieved, outperforming state-of-the-art approaches.

Reference

[1] Sanda Harabagiu and Andrew Hickl, Methods for Using Textual Entailment in Open-Domain Question Answering, In Proceedings of ACL 2006, 2006, pp 905-912.

[2] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Information fusion in the context of multi-document summarization. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park (MD), USA, 1999.

[3] Peter G. Anick and Suresh Tipirneni, 1999. The paraphrase search assistant: terminological feedback for iterative information seeking. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999.

[4] Kazutaka Shimada, Yasuto Seto, Mai Omura and Kohei Kurihara, KitAi: Textual Entailment Recognition System for NTCIR-10 RITE2, Proceedings of the 10th NTCIR Conference, June 18-21, 2013, Tokyo, Japan.

[5] Hen-Hsen Huang, Kai-Chun Chang and Hsin-Hsi Chen, Modeling Human Inference Process for Textual Entailment Recognition. The 51st Annual Meeting of the Association for Computational Linguistics - Short Papers (ACL Short Papers 2013), Sofia, Bulgaria, August 4-9, 2013.

[6] Zellig Harris. Distributional Structure. In: Katz, J. J. (ed.), The Philosophy of Linguistics. New York: Oxford University Press. pp. 26-47, 1985.

[7] Luisa Bentivogli, Peter Clark, Ido Dagan and Danilo Giampiccolo, The seventh pascal recognizing textual entailment challenge, Proceedings of TAC, 2011.

[8] Andrew Hickl and Jeremy Bensley. 2007. A discourse commitment-based framework for recognizing textual entailment. In Proceeding of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. pp 171–176.

[9] Peter Clark, Phil Harrison. 2009. An inference-based approach to recognizing entailment. In Proceeding of TAC, 2009 Workshop.

[10] Roy Bar-Haim, Jonathan Berant, Ido Dagan, Iddo Greental, Shachar Mirkin, Eyal Shnarch, and Idan Szpektor. 2009. Efficient semantic deduction and approximate matching over compact parse forests. In Proc. of the Text Analysis Conference 2008.

[11] Sebastian Padó, Marie-Catherine de Marneffe, Bill MacCartney, Anna N. Rafferty, Eric Yeh, and Christopher D. Manning. 2009. Deciding entailment and contradiction with stochastic and edit distance-based alignment. In Proceeding of TAC, 2008 Workshop.

[12] Mark Sammons, V.G. Vinod Vydiswaran, and Dan Roth. 2010. Ask not what textual entailment can do for you.... In Proceeding of the 48th Annual Meeting of the Association for Computational Linguistics. pp 1199–1208.

[13] Quang Do, Dan Roth, Mark Sammons, Yuancheng Tu and V.G.Vinod Vydiswaran. 2010. Robust, light-weight approaches to compute lexical similarity. Computer Science Research and Technical Reports, University of Illinois.

[14] Masaaki Tsuchida and Kai Ishikawa. IKOMA at TAC2011: A Method for Recognizing Textual Entailment using Lexical-level and Sentence Structure-level features. In Proc. Of Text Analysis Conference 2011.
[15] KenKen-ichi Yokote, Shohei Tanaka and Mitsuru Ishizuka. Effects of Using Simple Semantic Similarity on Textual Entailment Recognition. In Proceedings of the TAC 2011 Workshop.
[16] Yongmei Tan, Junyu Zeng, Xiaojie Wang, Eduard Hovy. 2011. BUPTTeam Participation at TAC2011 Recognizing Textual Entailment. In Proceedings of the TAC 2011 Workshop.