Lexical based two-way RTE System at RTE-5

Partha Pakray
Computer Science and Engineering Department,
Jadavpur University
Kolkata, India.

Sivaji Bandyopadhyay
Computer Science and Engineering Department,
Jadavpur University
Kolkata, India.

Alexander Gelbukh
Center for Computing Research,
National Polytechnic Institute,
Mexico City, Mexico

Abstract

The note describes the lexical based two-way Recognizing Textual Entailment (RTE) system developed at the Computer Science and Engineering Department, Jadavpur University, India. We participated in the two-way main task at RTE-5. The system is based on the composition of the following six lexical based RTE methods: WordNet based unigram match, bigram match, longest common sub-sequence, skip-gram, stemming and named entity matching. Each of these methods were applied on the development data to obtain two-way decisions. It was observed on the development data that final entailment decision on a text-hypothesis pair that is based on positive entailment decisions from at least two lexical based RTE methods was producing a better precision and recall figure. An accuracy figure of 58.17% was obtained on the test data. Ablation tests were performed for each of the six RTE methods and these are reported in the present note. The RTE task was based on three application settings: QA, IE and IR but this information was not taken into consideration during the system development. The relatively higher accuracy figures for the IR application setting obtained during the various tests suggest that identification of appropriate RTE methods based on the application settings might have improved the accuracy scores further.

1. Introduction

The main RTE-5 task is similar to the RTE-4 task, with the following changes:

- The average length of the RTE 5 texts were higher than RTE 4.
- Texts came from a variety of sources and were not edited from their source documents. Thus, RTE systems were required to handle real text that may include typographical errors and ungrammatical sentences.
- A development set was released.
- The textual entailment recognition task was based on only three application settings: QA, IE, and IR.

In addition to the main task (Textual Entailment *Recognition*), RTE-5 offered a new **Textual Entailment *Search*** pilot that is situated in the summarization application setting, where the task was to find all Texts in a set of documents that entail a given Hypothesis.

The main task in RTE-5 consists of two sub-tasks:

1. The three-way RTE task, where the system must decide whether:

   - T entails H - in which case the pair will be marked as ENTAILMENT
   - T contradicts H - in which case the pair will be marked as CONTRADICTION
   - The truth of H cannot be determined on the basis of T - in which case the pair will be marked as UNKNOWN

2. The two-way RTE task is to decide whether:

   - T entails H - in which case the pair will be marked as ENTAILMENT
   - T does not entail H - in which case the pair will be marked as NO ENTAILMENT

The RTE-5 development and the test set consisted of 600 text-hypothesis pairs each. The three applications – namely QA, IE, and IR – were considered as settings or contexts for the generation of each pair. 200 pairs were selected for each application in each data set.

The JU_CSE_TAC team from the Computer Science and Engineering Department, Jadavpur University participated in the two-way main RTE task. Section 2 describes the lexical based RTE system developed for taking part in the RTE-5 two-way main task. The various experiments carried out on the development and test data sets are described in Section 3 along with the results. As the final system was a combination of six lexical based RTE systems that are developed on six different knowledge sources, ablation tests were carried out to identify the relative importance of each of these different knowledge sources and the corresponding methods. The results of these ablation tests are reported in Section 4. The conclusions are drawn in Section 5.

2. System Description

This note describes a Lexical based approach for solving the Textual entailment problem. The final system is a combination of six different rule based system working on various lexical knowledge sources. The system computes the entailment decisions using each of these six lexical based RTE methods for each text-hypothesis pair. The final entailment decision is based on positive entailment decisions from at least two of these methods. The lexical based RTE methods are WordNet based unigram match, bigram match, longest common sub-sequence, skip-gram, stemming and named entity matching. Each of these methods are detailed in Section 2.2.

2.1. Pre-processing Task

The system accepts pairs of text snippets (text and hypothesis) at the input and gives a boolean value at the output: ENTAILMENT if the text entails the hypothesis and NO ENTAILMENT otherwise. An example text-hypothesis pair from the RTE-5 development set is shown in Figure 1.

```
<pair id="1" entailment="ENTAILMENT" task="QA">
			<t>The disappearance of York University chef Claudia Lawrence is now being treated as suspected murder, North Yorkshire Police said. However detectives said they had not found any proof that the 35-year-old, who went missing on 18 March, was dead. Her father Peter Lawrence made a direct appeal to his daughter to contact him five weeks after she disappeared. His plea came at a news conference held shortly after a £10,000 reward was offered to help find Miss Lawrence. Crimestoppers said the sum they were offering was &quot;significantly higher&quot; than usual because of public interest in the case.</t>
			<h>Claudia Lawrence is 35 years old.</h>
</pair>
```

Figure 1. RTE-5 development set text-hypothesis pair

In the development set, the entailment decisions were noted as a three-way task. This made it necessary to convert the three-way decisions in the development set to two-way decisions for the two-way main task. This conversion was carried out simply by merging “CONTRADICTION” and “UNKNOWN” decisions to “NO ENTAILMENT” decision. The English articles (a, an, the) were removed from both the text and hypothesis as these were occurring high in number. It was observed that escape characters like &quot; , &#133; , &#145; and &amp; are present in the text and the hypothesis parts and these were removed. All the above pre-processing methods were also applied on the test set.

2.2 Lexical based RTE methods

In this section the various lexical based RTE methods are described in detail.

1. **WordNet based Unigram Match:** In this method, the various unigrams in the hypothesis for each text-hypothesis pair are checked for their presence in the text. WordNet synsets are identified for each of the unmatched unigrams in the hypothesis. If any synset for the hypothesis unigram matches with any synset of a word in the text then the hypothesis unigram is considered as a WordNet based unigram match.

For example, let us consider the following text-hypothesis pair.

T: A whale that became stranded in the River Thames has died after a massive rescue attempt to save its life. The 18ft (5m) northern bottle-nosed whale was first spotted in the river on Friday and rescuers began an attempt to save it on Saturday morning. But the whale died at about 1900 GMT on Saturday as rescuers transported it on a barge towards deeper water in the Thames Estuary.
H: A whale died in the River Thames.

Here the common unigrams are whale, died, River, Thames.
If n1= common unigram or WordNet Synonyms between text and hypothesis and n2= number of unigram in Hypothesis,

Wordnet_Unigram_Match=n1/n2.

If the value of Wordnet_Unigram_Match is 0.75 or more, i.e., 75% or more unigrams in the hypothesis match either directly or through WordNet synonyms, then the text-hypothesis pair is considered as an entailment. The text-hypothesis pair is then assigned the value of 1 meaning entailment, otherwise, the pair is assigned a value of 0. The cut-off value for the Wordnet_Unigram_Match is based on experiments carried out on the RTE-5 main task development set.

2. **Bigram Match:** Each bigram in the hypothesis is searched for a match in the corresponding text part. The measure Bigram_Match is calculated as the fraction of the hypothesis bigrams that match in the corresponding text, i.e.,

Bigram_Match=(Total number of matched bigrams in a text-hypothesis pair /Number of hypothesis bigrams).

If the value of Bigram_Match is 0.5 or more, i.e., 50% or more bigrams in the hypothesis match in the corresponding text, then the text-hypothesis pair is considered as an entailment. The text-hypothesis pair is then assigned the value of 1 meaning entailment, otherwise, the pair is assigned a value of 0. The cut-off value for the Bigram_Match is based on experiments carried out on the RTE-5 main task development set.

3. **Longest Common Subsequence (LCS):** The Longest Common Subsequence of a text-hypothesis pair is the longest sequence of words which is common to both the text and hypothesis. LCS(T,H) estimates the similarity between text T and hypothesis H, as LCS_Match=*LCS*(*T,H*)/length of H.

If the value of LCS_Match is 0.8 or more, i.e., the length of the longest common subsequence between text T and hypothesis H is 80% or more of the length of the hypothesis , then the text-hypothesis pair is considered as an entailment. The text-hypothesis pair is then assigned the value of 1 meaning entailment, otherwise, the pair is assigned a value of 0. The cut-off value for the LCS_Match is based on experiments carried out on the RTE-5 main task development set.

4. **Skip-grams:** A skip-gram is any combination of n words in the order as they appear in a sentence, allowing arbitrary gaps. In the present work, only 1-skip-bigrams are considered where 1-skip-bigrams are bigrams with one word gap between two words in a sentence follow the order. The measure 1-skip_bigram_Match is defined as

1_skip_bigram_Match = *skip_gram*(*T,H*) / n,

where *skip_gram*(*T,H*) refers to the number of common 1-skip-bigrams (pair of words in sentence order with one word gap) found in T and H and *n* is the number of 1-skip-bigrams in the hypothesis H.

If the value of 1_skip_bigram_Match is 0.5 or more, then the text-hypothesis pair is considered as an entailment. The text-hypothesis pair is then assigned the value of 1 meaning entailment, otherwise, the pair is assigned a value of 0. The cut-off value for the 1_skip_bigram_Match is based on experiments carried out on the RTE-5 main task development set.

5. **Stemming**: Stemming is the process of reducing terms to their root form. So for example, the plural forms of a noun such as ‘boxes’ are transformed into ‘box’, and derivational endings with ‘ing’, ‘es’, ‘s’ and ‘ed’ are removed from verbs. Each word in the text and hypothesis pair is stemmed using the stemming function provided along with the WordNet 2.0.

If s1= number of common stemmed unigrams between text and hypothesis and s2= number of stemmed unigrams in Hypothesis, then the measure Stemming_match is defined as

Stemming_Match=s1/s2

If the value of Stemming_Match is 0.7 or more, i.e., 70% or more stemmed unigrams in the hypothesis match in the stemmed text, then the text-hypothesis pair is considered as an entailment. The text-hypothesis pair is assigned the value of 1 meaning entailment, otherwise, the pair is assigned a value of 0. The cut-off value for the Stemming_Match is based on experiments carried out on the RTE-5 main task development set.

6. **Named Entity Match**: It is based on the detection and matching of Named Entities (Nes) in the text-hypothesis pair. Once the NEs of the hypothesis and the text have been detected, the next step is to determine the number of Nes in the hypothesis that match in the corresponding text. The measure NE_Match is defined as

NE_Match=number of common NEs between text and hypothesis/Number of NE in Hypothesis.

If the value of NE_Match is 0.5 or more, i.e., 50% or more NEs in the hypothesis match in the text, then the text-hypothesis pair is considered as an entailment. The text-hypothesis pair is assigned the value of 1 meaning entailment, otherwise, the pair is assigned a value of 0. The cut-off value for the NE_Match is based on experiments carried out on the RTE-5 main task development set.

For named entity recognition, the RASP Parser (Briscoe et al., 2006) nertag component has been used. The nertag component is a rule-based named entity recognizer which recognizes and marks up the following kinds of named entity: numex (sums of money and percentages), timex (dates and times) and enamex (persons, organizations and locations).

WordNet [Fellbaum, 1998] is one of most important resource for lexical analysis. The WordNet 2.0 has been used for WordNet based unigram match and stemming step. API for WordNet Searching (JAWS) [Brett Spell] is an API that provides Java applications with the ability to retrieve data from the WordNet database.

3. Experiments on the Development and the Test data and the results

The RTE-5 main task development set was used to train the various RTE methods to identify the cut-off values for the various measures towards entailment decision. The RTE-5 development set consisted of 600 text-hypothesis pairs. The RTE-5 main task test set consisted of 600 text-hypothesis pair.

In our lexical based textual entailment system, each method was run separately on the RTE-5 development set and two-way entailment (ENTAILMENT or NO ENTAILMENT) decisions were obtained for each text-hypothesis pair. Experiments were carried out to measure the performance of the final RTE system based on whether the final entailment decision is based on positive results from 2, 3 or 4 methods. The results are shown in Table 1. It is observed that the precision and recall measures of the final RTE system is best when final entailment decision is based on positive results from at least two methods.

Development Set	**2 methods**	**3 methods**	**4 methods**
The system number of entailment	381	235	121
The system compared with gold number of entailment	223	134	73
In Gold data number of entailment	300	300	300
Precision	0.58	0.57	0.60
Recall	0.74	0.44	0.24

Table 1: Textual Entailment System Results in RTE 5 Development Set in the two-way entailment classification

The RTE system was run on test data and two runs were submitted. The first run was not ranked and for each text-hypothesis pair, ENTAILMENT or NO ENTAILMENT decisions were noted. The second run ranked with all the ENTAILMENT decisions placed at the first part of the ranked list and the NO ENTAILMENT decisions placed later. The ranking decision was based on the number of methods that generated positive entailment decisions for the text-hypothesis pair. The evaluation results as identified by the organizers for the two runs are mentioned in Table 2.

	Accuracy		
	Run 1 (NO RANK)	Run 2 (RANK)	
	accuracy2w	accuracy2w	Average precision
QA	0.55	0.55	0.5259
IE	0.51	0.51	0.4854
IR	0.685	0.685	0.6966
Overall	0.5817	0.5817	0.5508

Table 2: Result for the test set : Two way Run 1 (NO RANK), Run 2 (RANK)

As we can see in the run, different accuracy values are obtained depending on the task. The worst result is obtained in the IE task.

4. **Ablations test and results**

An ablation test consists of removing one module at a time from a system, and rerunning the system on the test set with the other modules, except the one tested. Comparing the results to those obtained by the system as a whole, it is possible to assess the practical contribution given by each single module.

In order to better understand the relevance of the knowledge resources used by RTE systems, and evaluate the contribution of each of them to the systems' performances, ablation tests for major knowledge resources required for those systems.

The two-way RTE task decision between the text and the hypothesis is based on the following methods: WordNet based Unigram match, Bigram match, Longest Common Subsequence, Skip bigram, Unigram match after stemming and Named Entity match.

In abl-1, we ablated WordNet based Unigram match.
In abl-2, we ablated Named Entity match.
In abl-3, we ablated Skip bigram match.
In abl-4, we ablated Bigram match.
In abl-5, we ablated Longest Common Subsequence.
In abl-6, we ablated stemming.

	accuracy2w					
	abl-1	**abl-2**	**abl-3**	**abl-4**	**abl-5**	**abl-6**
QA	0.545	0.545	0.55	0.55	0.54	0.54
IE	0.525	0.515	0.51	0.51	0.535	0.515
IR	0.665	0.685	0.685	0.685	0.685	0.705
Overall	0.5783	0.5817	0.5817	0.5817	0.5867	0.5867

Table 3: Ablation Test Results with accuracy2w

	rel_accuracy2w					
	abl-1	**abl-2**	**abl-3**	**abl-4**	**abl-5**	**abl-6**
QA	-0.005	-0.005	0	0	-0.01	-0.01
IE	0.015	0.005	0	0	0.025	0.005
IR	-0.02	0	0	0	0	0.02
Overall	-0.0034	0	0	0	0.005	0.005

Table 4: Ablation Test Results with rel_accuracy2w

5. **Conclusions**

Results show that a lexical-based approach is not enough to tackle appropriately the textual entailment problem. Experiments have been started for a syntax based RTE task using dependency parser. In the present task, the final RTE system has been optimized for the ENTAILMENT decision using the development set while the optimization for the NO ENTAILMENT decision would have been an interesting experiment to look into. The role of the application setting for the RTE task has also not been looked into. This needs to be experimented in future. Finally, the two way task has to be upgraded to the three way task.

References:

1. Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 1–9, Prague, June. Association for Computational Linguistics.
2. Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Bill Dolan. 2008. The Fourth PASCAL Recognizing Textual Entailment Challenge. Proceedings of the First Text Analysis Conference (TAC 2008), November 17-19. Gaithersburg, Maryland, USA
3. Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Mass.
4. E. Briscoe, J. Carroll, and R. Watson. 2006. The Second Release of the RASP System. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions.
5. Zornitsa Kozareva, Sonia Vazquez and Andres Montoyo. 2006. Adaptation of a Machine-learning Textual Entailment System to a Multilingual Answer Validation Exercise, CLEF Working Notes.
6. Java API for WordNet Searching (JAWS), http://lyle.smu.edu/~tspell/jaws/index.html