Detecting Textual Entailment with Conditions on Directional Text Relatedness Scores Revisited

Alpár Perini [palpar at gmail.com]

1 Introduction

The system described below is based on the one described in paper [Per] that uses directional text relatedness conditions for detecting textual entailment between sentences. Some other conditions have been experimented with that were generated automatically using a variant of Genetic Programming. The sentences were searched regardless of the context in which they appear, as the formulas used were not directly taking that into account. The word relatedness score required by the formula uses not only identity and synonymy, but almost all the WordNet relations. The part of speech tagging was made using the latest version of the Stanford POS Tagger. We show the results that we have obtained using our implementations for the RTE-2010 test data and the ablation testing performed after.

2 Overview of the DirRelCond2 System

The theoretical background used by the DirRelCond2 system was presented in the paper [Per] describing the earlier DirRelCond system for detecting textual entailment.

Recall here the formula proposed in [TSMM09] for calculating the *text relatedness score*:

$$rel(T,H)_T = \frac{\sum_{pos} \sum_{T_i \in WS^T_{pos}} (maxRel(T_i) \times idf(T_i))}{\sum_{pos} \sum_{T_i \in WS^T_{pos}} idf(T_i)} \quad (1)$$

3 Inside the DirRelCond2 System – The Conditions

In this section we describe the component of our system, which uses (directional) conditions on the above mentioned relatedness scores for discovering

entailment relations.

Condition (4) from [Per] was empirically tuned for the RTE-2010 development dataset:

$$rel(T,H)_H > rel(T,H)_T + 0.6 \tag{2}$$

In addition, we have experimented with other, more complex conditions for detecting entailment. These conditions were earlier generated using Gene Expression Programming (GEP) [Fer01, Olt09], using the 2009 development dataset as reference.

Two of the individuals we have experimented with, and not used in the previous DirRelCond system, were representing heuristics of the form:

$$rel(T,H)_T < 0.4527 \times rel(T,H)_H^3 \tag{3}$$

$$rel(T,H)_T + 1.15 < rel(T,H)_H^2 + rel(T,H)_H \tag{4}$$

4 Experimental Results

The DirRelCond2 application was developed in Java for recognizing textual entailment using the proposed conditions. The new component that was introduced was for parsing all the input data given in the particular format and constructing an object hierachy of it. This made it possible to form hypothesis and text pairs as it was accepted by the earlier DirRelCond system, which had to be slightly modified to use the new entailment conditions. The system takes into account only these two sentences when deciding on the trueth value of the entailment, ingnoring the context of the text that they are part of, as it was the case in previous challanges.

A part of speech tagger was needed in order to distinguish the open class words. We used the latest Stanford POS tagger implemented in Java [sta10] for finding the sets of open-class words. For looking up words and word relations, we used WordNet 2.1 [Fel98], accessed through the Java interface provided by JWordNet [Fei08].

We worked with all the possible senses for T_i with the given *pos*. The current implementation simplifies the relatedness formula by considering $idf(w)$ to be always 1 and hence the importance of a word w with respect to some documents is neglected. A potential improvement would be to take into account the entire set of sentences found in the development dataset and use that to obtain this value.

Our application participated at the RTE-2010 challenge, therefore it was run several times against the development and testing datasets. The results of the accuracies obtained are summarized in Table 1 below.

System	*Precision(%)*	*Recall(%)*
Run 1 (2)	38.99	41.80
Run 2 (3)	52.38	15.13
Run 3 (4)	61.76	17.78

Table 1: Comparison of RTE-2010 precisions and recalls obtained by our DirRelCond2 system for the testsets.

The precision results show that condition (4) performed better than the other conditions for the test set. Condition (3) and mainly condition (2) did not scale well for newly seen data. However, condition (2) obtained the best recall measure, while the others were significantly worse. This means that if we are interested in discovering as many potential entailments as possible, condition (2) is better, while if we want a greater certainty for the entailment to hold, then (4) is a compromise solution. Overall the results are acceptable if we take into account that no sentence context information was used for producing the results.

The last stage of the contest required to perform ablation tests. As opposed to last year's approach, where we considered not breaking the unity of the formula, we attempted to remove two crucial parts of the system, one after the other. Condition (4) was the one being tested, since that produced the best results. The first experiment, called abl_1 had WordNet removed. This way only basic word comparison was used instead of word relations. The word relatedness score was a 1 only in case of identical words, otherwise it was 0. The second experiment, called abl_2 had in addition to abl_1 the part of speech tagger removed as well and therefore all words considered as having the same POS, forming one big set.

Components	*Precision(%)*	*RelPrec(%)*	*Recall(%)*	*RelRec(%)*
abl_1	70.00	-8.24	11.11	+6.67
abl_2	75.97	-14.21	12.38	+5.40

Table 2: Comparison of RTE-2010 ablation testing precisions and recalls

After analyzing the ablation testing results, we can say that the precision

of the system obtained after removing some components was better, however the recall had diminished significantly. This is partly expected, because the system was not running according to the formulas. This way the condition became too restrictive, therefore the few conditions that were detected as entailment were more likely to be indeed true.

References

[Fei08] Ingo Feinerer. *wordnet: WordNet Interface*, 2008. R package version 0.1-3.

[Fel98] Christiane Fellbaum. *WordNet: An Electornic Lexical Database*. Bradford Books, 1998.

[Fer01] Candida Ferreira. Gene expression programming: a new adaptive algorithm for solving problems. *ArXiv Computer Science e-prints*, February 2001.

[Olt09] Mihai Oltean. Genetic Programming – Automatic Source Code Generation course. Technical report, Babes-Bolyai University, 2009.

[Per10] Alpar Perini. Detecting textual entailment with conditions on directional text relatedness scores. In *Proceedings of the Second Text Analysis Conference (TAC 2009)*, pages 1–10. NIST, 2010.

[sta10] Stanford POS tagger, Jun 2010.

[TSMM09] Doina Tatar, Gabriela Serban, A. Mihis, and Rada Mihalcea. Textual entailment as a directional relation. *Journal of Research and Practice in Information Technology*, 41(1):17–28, 2009.