A Knowledge-based Textual Entailment Approach applied to the QA Answer Validation at CLEF 2006

Ó. Ferrández, R. M. Terol, R. Muñoz, P. Martínez-Barco and M. Palomar
Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{ofe,rafamt,rafael,patricio,mpalomar}@dlsi.ua.es

Abstract

The Answer Validation Exercise (AVE) is a pilot track within the Cross-Language Evaluation Forum (CLEF) 2006. The AVE competition provides an evaluation framework for answer validations in Question Answering (QA). In our participation in AVE, we propose a system that has been initially used for other task as Recognising Textual Entailment (RTE). The aim of our participation is to evaluate the improvement our system brings to QA. Moreover, due to the fact that these two task (AVE and RTE) have the same main idea, which is to find semantic implications between two fragments of text, our system has been able to be directly applied to the AVE competition. Our system is based on the representation of the texts by means of logic forms and the computation of semantic comparison between them. This comparison is carried out using two different approaches. The first one managed by a deeper study of the WordNet relations, and the second uses the measure defined by Lin in order to compute the semantic similarity between the logic form predicates. Moreover, we have also designed a voting strategy between our system and the MLEnt system, also presented by the University of Alicante, with the aim of obtaining a joint execution of the two systems developed at the University of Alicante. Although the results obtained have not been very high, we consider that they are quite promising and this supports the fact that there is still a lot of work on researching in any kind of textual entailment.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Algorithms, Semantic Similarity, Experimentation, Measurement, Performance

Keywords

Question Answering, Answer Validation, Textual Entailment, WordNet, Semantic Relations

1 Introduction

The Answer Validation Exercise (AVE) is a pilot track within the Cross-Language Evaluation Forum (CLEF) 2006. The aim of AVE is to provide an evaluation framework for answer validations in Question Answering (QA) systems. This automatic Answer Validation would be useful for improving the performance of QA systems, helping humans in the assessment of QA systems

output, improving QA systems self-score, developing better criteria for collaborative QA systems, etc.

The organizers of AVE took an answer plus a snippet given by a QA system, and they built a hypothesis turning the question plus the answer into an affirmative form. If the given text (a snippet or a document) semantically entails this hypothesis, then the answer is expected to be correct. They provided pairs text-hypothesis for the participants which have to determine if the entailment holds. The final purpose is quite similar to the purpose of other challenges as the PASCAL Recognising Textual Entailment [1].

In a nutshell, the participant systems must emulate human assessments of QA responses and decide whether an answer is correct or not according to a given snippet.

In our participation in AVE, we want to evaluate the positive impact that our system can produce in the context of QA. Initially, our system was developed for Recognising Textual Entailment (RTE) by means of snippets in English language. However, due to the fact that these two task (AVE and RTE) have the same main idea, which is to find semantic implications between two fragments of text, our system has been able to be directly applied to the AVE competition.

The rest of this paper is organized as follows. The following section presents the description of our system and its components. Section 3 illustrates the experiments carried out and the results obtained. Finally, section 4 wraps up the paper with some conclusions and future work proposals.

2 System Description

As we have mentioned in the previous section, the system that we describe here has already been used to solve Textual Entailment. A detailed description of our system is depicted in [6]. In this paper, we only make a brief overview of the components that our system is composed of, and how these components work in order to find an entailment relation between two text fragments.

Our system has two main components: (i) the first one obtains the logic forms associated to each text; and (ii) the second computes the semantic similarity between the aforementioned logic forms. These components will be detailed in the followings paragraphs.

The process our system follows is the following:

1. It obtains the logic forms from the two given texts.
2. It computes the semantic similarity between the generated logic forms. This step will provide a semantic weight that will determine a true or false entailment.
3. It compares the semantic weight obtained in the previous step to an empiric threshold acquired from the development corpus.

2.1 Derivation of the Logic Forms

A logic form can be defined as a set of predicates related among them which have been inferred from a sentence. The aim of using logic forms is to simplify the sentence treatment process.

In our approach, we use a format for representing logic forms similar to the format of the lexical resource called Logic Form Transformation of eXtended WordNet (LFT) [2]. And the process to infer the logic form associated of a sentence is through applying NLP rules to the dependency relationship of the words. Thus, the first step is to obtain the dependency relationships between the words of the sentence. We use MINIPAR [4], a broad-coverage parser, in order to obtain these dependency relationships.

Once the dependency relationships have been acquired, the next step is the analysis of these dependencies by means of several NLP rules that transform the dependency tree into its logic form associated.

To sum up, the derivation of logic forms consists of a compositional process that starts in the leaves of the dependency tree, continues through the ramifications and ends in the root of the dependency tree.

2.2 Computation of Similarity Measures

The main idea of this component is that the verbs generally govern the meaning of sentences. For this reason, this method is initially focused on analysing semantic relations between the verbs of the two logic forms derived from the text and the hypothesis respectively. And secondly, if there is a relation between the verbs, then the method will analyse the similarity relations between all predicates depending on the two verbs. In the case of there is not semantic relations between the verbs, this method will not analyse any more logic form predicate.

In order to obtain the similarity between the predicates of the logic forms, two approaches have been implemented:

- **Based on WordNet relations:** we determine if two predicates are related through the composition of the WordNet relationships. We consider hyponymy, entailment and synonymy WordNet relations between the predicates from the text to the hypothesis. And, if there is a path which connects these two predicates, we conclude that these predicates are semantically related with a specific weight. The length of the path that relates the two different predicates must be lower or equal than 4. Each WordNet relation has assigned a weight, and the weight of the path is calculated as the product of the weights associated to the relations connecting the two predicates.

- **Based on Lin's measure [5]:** in this case, the semantic similarities were computed using Lin's similarity measure as is implemented in WordNet::Similarity[1] [7]. Lin's similarity measure augments the information content of the least common subsumer (LCS[2]) of the two concepts with the sum of the information content of the concepts themselves. The Lin's measure scales the information content of the LCS by this sum.

A Word Sense Disambiguation module was not employed in deriving the WordNet relations between any two predicates. Only the first 50% of the WordNet senses were taken into account.

2.3 UA-voting

As the University of Alicante has two systems based on different techniques which solve the recognition of Textual Entailment. We want to evaluate each system in the very recent AVE task individually as well as check how a combination of these two systems could improve the results. The systems involved in this experiment were: our system explained in this paper and the system presented by Kozareva et al. [3], called MLEnt.

For the purpose of testing this combination, we sent a run combining the outputs of the two systems. This combination was carried out for English language and we merged the outputs with the simplest method to combine systems, a voting strategy.

We composed the final output by means of three different outputs. The final result suggested by our voting strategy must coincide with two individual outputs. The three considered outputs were: our output with the module of semantic similarity using Lin's measure and two outputs provided by MLEnt regarding two different experiments about skip-grams and the longest common subsequence technique[3].

3 Results and Discussion

For the development and test of our system, we used the corpus provided by the AVE organizers. The corpora consist of a set of pair text-hypothesis built semi-automatically from QA@CLEF 2006 responses and the results returned by the participants will be evaluated against the QA human assessments.

[1] http://www.d.umn.edu/~tpederse/similarity.html
[2] LCS is the most specific concept that two concepts share as an ancestor
[3] For further details see [3]

The development corpus for English has around 2870 pairs test-hypothesis, but only 168 are revised manually. We only used the revised pairs in order to adjust our system for the AVE task. The test data contains 2088 pairs, and all the results obtained are shown in Table 1. In this table, we illustrate the results achieved by our two semantic similarity approaches individually (see section 2.2) and the results obtained regarding UA-voting experiment (see section 2.3).

Development data	Precision YES pairs	Recall YES pairs	F-measure
WNrelations	0.2368	0.75	0.36
Lin	0.2265	0.8055	0.3536
Test data	**Precision YES pairs**	**Recall YES pairs**	**F-measure**
WNrelations	0.2072	0.5116	0.2949
Lin	0.1981	0.6884	0.3077
UA-voting	0.2054	0.6047	0.3066

Table 1: *AVE 2006 officials results for English language*

As we can observe in Table 1, all the results are quite similar with respect to F-measure. Using the approach based on Lin's semantic similarity measure our system achieved better recall than using the approach about WordNet relations. However, these differences are insignificant to decide what approach works better for the AVE task.

The run corresponding to the combination of the two systems developed at the University of Alicante did not achieve the expected results. These results prove that we have to investigate other ways in order to combine the outputs of the systems, other voting strategies or, perhaps to join the two different technologies of each system in order to create only one system.

4 Conclusions and Future Work

In this paper, we have presented a system based on the representation of the texts by means of logic forms and the computation of semantic comparison between them. This comparison is carried out using two different approaches. The first one managed by a deeper study of the WordNet relations between the predicates of the text and the hypothesis, and the second uses the measure defined by Lin [5] in order to compute the semantic similarity between the logic form predicates.

This system has already been applied to Recognising Textual Entailment (see [6]), but in this case the aim of applying it to the AVE task was to check the improvement our system brings to QA. Moreover, we also present in this paper a voting strategy combining the two systems developed at the University of Alicante: our system and the system presented by Kozareva et al. [3] for the AVE task.

The results obtained have not been very high, but quite promising. However, we want to attach great importance to the fact that, in the RTE-2 Challenge [1] our system achieved 60% in average precision, but for the AVE task the result has decreased dramatically. This supports the claim that research in any kind of textual entailment is still at the very first steps and so, there is a long way to go.

As a future work, We want to investigate in depth the corpus provided by AVE and find the cases that our system fails and why. Possibly, in order to solve these deficiencies of our system, we need to improve our method by investigating in more detail the syntactic trees of the text and the hypothesis and how the addition of other NLP tools such as a Named Entity Recognizer could help in detecting entailment between two segments of text. Finally, with this kind of knowledge we will be able to integrate our system within a module performing answer validation for QA.

Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01.

References

[1] Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The Second PASCAL Recognising Textual Entailment Challenge. *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge, RTE-05*, pages 1–9, 2005.

[2] S. Harabagiu, G.A. Miller, and D.I. Moldovan. WordNet 2 - A Morphologically and Semantically Enhanced Resource. In *Proceedings of ACL-SIGLEX99: Standardizing Lexical Resources*, pages 1–8, Maryland, June 1999.

[3] Zornitsa Kozareva and Andrés Montoyo. MLEnt: The Machine Learning Entailment System of the University of Alicante. *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge, RTE-05*, pages 16–21, 2005.

[4] D. Lin. Dependency-based evaluation of minipar. In *Workshop on the Evaluation of Parsing Systems*, pages 17–20, Southampton, UK, April 2005.

[5] Dekang Lin. An Information-Theoretic Definition of Similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[6] Óscar Ferrández, R. M. Terol, Rafael Mu noz, Patricio Martínez-Barco, and Manuel Palomar. An Apporach based on Logic Forms and WordNet relationships to Textual Entailment Preformance. *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge, RTE-05*, pages 22–26, 2005.

[7] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet::Similarity - Measuring the Relatedness of Concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, San Jose, CA, July 2004.