Alicante University at TAC 2009: Experiments in RTE

Óscar Ferrández, Rafael Muñoz, and Manuel Palomar
Natural Language Processing and Information Systems Group
Department of Computing Languages and Systems
University of Alicante
{ofe, rafael, mpalomar}@dlsi.ua.es

Abstract

This paper discusses the participation of the University of Alicante in the RTE track at TAC 2009. The proposed system faces the entailment recognition by computing shallow lexical deductions and richer inferences based on semantics. Specifically on WordNet, detection of negation terms, named entity recognition, verbs implications and frame semantic analysis. Although the system was designed to deal with 2-way entailment classifications, we also wanted to test its behaviour when tackling the 3-way task. The results achieved overcame the median reached for all participants, however, when processing the ablation tests, which are new measures to evaluate the participants introduced this year, they revealed that despite the effort made to apply semantic knowledge, slight improvements were accomplished by the sophisticated part of the system, which encourages us even more to go on to further research in this line.

1 Introduction

The Recognizing Textual Entailment (RTE) track at the Text Analysis Conference (TAC) 2009 aims to evaluate the capabilities of the participants in recognizing when the meaning of one piece of text (the *text* or T) entails another (the *hypothesis* or H), which has been previously defined as an entailment relation between two snippets [7]. In addition, the RTE track at TAC continued the efforts of the PASCAL RTE Challenges [11] as well as the previous RTE-TAC 2008 edition [10].

In our participation, we present a system that integrates several inferences from different knowledge sources. The foundation of the system is mainly based on lexical deductions, afterwards several modules have been added to the system in order to compute more sophisticated deductions (e.g. WordNet relations, Named Entities (NEs) correspondences and verbs relations, FrameNet deductions, etc.).

This paper is organized as follows: Section 2 is intended to provide a detailed description of our textual entailment system carefully explaining each inference used, Section 3 presents the experiments carried out together with a discussion about the results obtained, and Section 4 contains conclusions regarding the work done.

2 Discovering Entailment Relations: our RTE System Description

The idea behind the proposed system is to develop an inference-based approach to solve entailment relations. Its principal goal is to tackle the entailment phenomena from different angles, thus the inferences performed (which are supported by lexico-semantic resources) will enable us to determine when an entailment relation takes place.

The system work-flow starts by taking the pair text-hypothesis as input and, afterwards, the developed inferences are responsible for extracting a set of features that will be passed to a machine learning algorithm. Specifically, we use the Weka's Support Vector Machine algorithm implementation [25], which has been stated in previous works (including ours) and achieves good performance in the task of recognizing entailment relations [4, 22, 5, 1]. All the system's inferences will be profoundly explained in the next subsections.

2.1 Shallow Knowledge: String-based Similarities

In previous RTE challenges, it has been demonstrated that string-based overlappings, whilst practically knowledge-less techniques, obtain promising results and in many cases are the base of lots of textual entailment systems. Hence, as in our previous participations, our system implements a module focused on such techniques.

This module carries out the computation of several string-based measures over the lemmata belonging to the two snippets (without considering stopwords). In order to obtain the lemmata and the part-of-speech information the Freeling toolkit was used [2], and as a result each measure produces a score that shows the similarity degree between the target snippets. These scores will serve as features for our machine learning algorithm.

The set of string-based measures comprises a binary matching, the Levenshtein distance [15], the Smith-Waterman algorithm [23], the Needleman-Wunsch algorithm [18], the Jaro distance [12], a matching between consecutive subsequences, the Cosine similarity, the Soundex distance[1], and the Inverse Document Frequency (IDF) specificity computed as in [24] and using the the LA Times 94 and Glasgow Herald 95 collections (169,477 documents) from the Cross-Language Evaluation Forum[2] (CLEF). Further details about the measures can be found in our previous paper [9].

[1] http://en.wikipedia.org/wiki/Soundex

[2] http://www.clef-campaign.org/

2.2 Richer Knowledge: Lexical-Semantic Deductions

Obviously, such a semantic problem cannot be completely and properly solved without integrating semantic knowledge into the system. Thus, we decided to check if more sophisticated knowledge such as that provided by resources like WordNet [17] and FrameNet [3] could help in detecting entailments.

2.2.1 Measuring WordNet-based Similarities

This analysis derives a score indicating the similarity degree focused on the semantic relation between words encoded in WordNet. In contrast to other approaches, in our case we also consider within the final score the words that are not found in WordNet by computing the Smith-Waterman algorithm between them. It allows us to take into account entities, which while not appearing in WordNet, are of paramount importance to determine the entailment relation.

For our inferences, we have used the Java WordNet Similarity Library [20] and the WordNet::Similarity::Tool [19], performing an accumulated score obtained by the sum of the maximum similarities achieved between the lemmata of the hypothesis with regards to the lemmata of the text. We consider four WordNet-based measures in order to obtain the maximum similarity: `Resnik` [21], `Lin` [16], `Jiang & Conrath` [13] and `Pirro & Seco` [20], and they are processed over the nouns, verbs, adjectives, adverbs and considering all the aforementioned grammatical categories,[3] obtaining a similarity score for each grammatical group, and consequently the corresponding learning features.

Furthermore, we also exploited another way to give more relevance to the semantic connections found in WordNet. It consisted of weighting the similarities according to the IDF values. So, each maximum similarity corresponding to each hypothesis' lemma is weighted by its IDF value. The final score considering the IDF information is also passed as a system feature.

2.2.2 Features Related to Negation

Regarding negation, we have implemented two inferences that will potentially help the system to support the entailment decision:

- The antonymy WordNet and VerbOcean [6] relation. A feature was created indicating whether verbs appear in the text having an antonymy relation with any verb in the hypothesis.[4]

- The general polarity regarding negative terms. We elaborated a basic list of negative terms to be used to deduce a feature showing the polarity of the pair text-hypothesis according to the number of occurrences of such terms.

[3] We used the Java WordNet Similarity Library for processing nouns and the WordNet::Similarity::Tool for the rest of categories.

[4] For antonymy relations we took the most frequent sense.

2.2.3 Inferences based on Named Entities

Based on the detection, absence and presence of NEs, we developed some inferences that show a certain degree of entailment focused on NEs.

The idea is somewhat simple, and consists of finding correspondences between the entities appearing in the texts. These correspondences will be established by a partial entity matching as well as an acronyms' deduction (i.e. "Gabriel García Márquez" ⇔ "García Márquez", "IBM" ⇔ "International Business Machines").

Therefore, two features were added to the system regarding NEs:

- A binary feature showing if all hypothesis' entities have at least one correspondence with the NEs in the text.
- A normalized value showing how many hypothesis' entities have correspondences.

Moreover, this knowledge was also integrated as a prior constraint that will discard those pairs (*entailment* = *no*) where there is at least one entity in the hypothesis without correspondence with regards to the text's entities. The computation of this constraint will be optional, and we will evaluate its impact in section 3.

2.2.4 Inferences based on Verbs Relations

Similar to the inferences based on NEs, we wanted to find out relations between the main verbs present in the entailment pair.

In this event, a correspondence between two verbs is found if: (i) they share the same lemma or they are synonyms regarding WordNet; (ii) they belong to the same VerbNet [14] class or a subclass of their classes; and (iii) there is a relations in VerbOcean that connects them.[5]

As for the NEs, two features were added to the system as well as a constraint prior to the computation of the inferences. In this case all based on relations between verbs.

2.2.5 Frame Semantic Analysis

Applying frame semantic analysis we wanted to check if the robustness offered by resources such as FrameNet [3] can help in determining the entailment. Our aim is to obtain similarity factors based on FrameNet that denote entailment relations, such factors will be added to the system as learning features.

To do this, the first step was to annotate the texts with frames and frame elements (roles) using the Shamaneser tool [8]. Once this step was complete, we proceeded to develop a *frame elements overlapping* that shows how many frame

[5] We do not take into account the antonymy and enablement relations because the antonymy relation was already considered in the negation features, and the enablement one neither involves transaction nor symmetry.

elements from the frames detected in both T and H share similar or lexically related instantiations. To compare frame element instantiations, we used the Levenshtein distance (with a similarity threshold higher or equal to 80%) as well as the WordNet synonym and hyponym relations from T's frame elements to H's frame elements. Therefore, two frame element instantiations are similar if they have the same lemma, their Levenshtein distance is higher or equal to 80% or the T's instantiation is a synonym or hyponym of the H's instantiation. Note that this sort of overlapping is more robust than those based on lexical transformations presented in section 2.1.

However, there are cases where even though different frames appear in T and H, they are connect by the semantic relations encoded in FrameNet, and such situations can discover positive entailments. Hence, we developed a Frame-to-Frame similarity procedure that obtains a similarity score focused on finding a semantic path connecting two frames through the FrameNet relations. This score quantifies how much two frames are alike based on the information contained in FrameNet.

To find the connection path, we explore a maximum depth of 5, because in our experiments longer paths reported insignificant semantic values. The final similarity score between two frames ($F \rightarrow F'$) is obtained as shown in the equation 1:

$$\begin{aligned} F2F_{sim} = FW(F) \, * \, W_{R1} \, * \, FW(f_1) \, * \, W_{R2} \, * \, RW(R_1, R_2) \, * \, ... \\ * \, RW(R_{n-1}, R_n) \, * \, W_{Rn} \, * \, FW(F') \end{aligned} \quad (1)$$

where $FW(x)$ measures the generality of each frame within the path by the inverse function of the number of children the frame has according to the relations involved in the connection path. $W(Rx)$ is the weight associated with the specific Rx relation, these weights were established heuristically considering the significance that each relation has in the FrameNet hierarchy and they are shown in Table 1. Finally, $RW(Rx, Ry)$ is the weight assigned when Rx precedes Ry in the path. To establish these weights, the idea was that when the same or similar relation is following in the path (e.g. from *Inherits_from* to *Inherits_from*) the weight is equal to one, when it goes down (e.g. from *Inherits_from* to *Inherited_by*) the weight is the same for the parent relation (see Table 1), and when it goes up (e.g. from *Inherited_by* to *Inherits_from*) the weight is equal to the parent relation weight divided by 2. The underlying intuition to take this decision is that from child nodes to parent nodes the loss of semantic information is higher than from parent nodes to child ones. For cross-relation references: *Inheritance*, *Perspective on* and *Using* are considered as similar relations in order to establish these weights, as well as *Causative* and *Inchoative*. The rest of the cross-relation reference values, although seldom appearing, are established to one leaving the similarity decision to the other weights.

Finally, the similarity score will be integrated as a feature for the learning and testing phase of the system.

Relation	**Parent (FROM)**	**Child (BY)**
Membership	1	
Inheritance	0.8	0.7
Perspective on	0.8	0.7
SubFrame	0.6	0.5
Precedes	0.7	0.7
Causative	0.5	0.4
Inchoative	0.5	0.4
Using	0.7	0.6
See also	0.5	0.4

Table 1: Frame-to-Frame: FrameNet relation weights.

3 Experiments, Results and Discussion

Table 2 shows the results obtained for both the development and test corpus, as well as for every experiment carried out. Although the system was not designed in principle to deal with 3-way entailment classification, due to the fact that a 3-way training corpus was also released we wanted to check our system behaviour in tagging three different kinds of entailments. Table 2 illustrates these results, as well.

Table 2: Results obtained for the RTE-TAC 2009 2-ways and 3-ways tasks.

RTE 2-ways task					
Run	Dev. corpus	Test corpus			
	overall	overall	QA	IE	IR
$ALL_{inferences}$	0.6533	0.6283	0.565	0.525	0.795
$ENT_{constraint}$	–	**0.6317**	0.57	0.525	0.8
$BOTH_{contraints}$	–	0.62	0.565	0.5	0.795
$Participants_{HIGH}$	0.7350				
$Participants_{MEDIAN}$	0.6117				
$Participants_{LOW}$	0.5000				
RTE 3-ways task					
Run	Dev. corpus	Test corpus			
	overall	overall	QA	IE	IR
$ALL_{inferences}$	0.6083	**0.6**	0.575	0.5	0.76
$ENT_{constraint}$	–	0.51	0.455	0.48	0.598
$BOTH_{contraints}$	–	0.4717	0.43	0.425	0.56
$Participants_{HIGH}$	0.6833				
$Participants_{MEDIAN}$	0.5200				
$Participants_{LOW}$	0.4383				

During the training phase, we developed a feature selection procedure directed by the information gain each feature provides to the classification problem. After that, some lexical features were discarded as they overlap each other resulting in noisy features. Also, the WorNet-based features applied just to adverbs and adjectives were discarded for the final set of features.

We carried out three experiments: (1) considering all the best features ($ALL_{inferences}$); (2) applying the constraint based on NEs correspondences prior to the computation of the inferences ($ENT_{constraint}$), which avoided processing 19% of the test corpus; and (3) also considering the constraint about H-to-T verbs relations ($BOTH_{contraints}$), which reduced the test corpus processing by 26%.

As shown in Table 2, the system does not perform very well when it faces the 3-way task. Specifically, the constraints produce a negative effect in the 3-way task because each pair that does not surpass the constraints is tagged as "CONTRADICTION" and, although it works properly for the 2-way classification, for the 3-way one it fails in most cases since the correct classification is "UNKNOWN". Nevertheless, considering the system was neither designed nor prepared to recognize three entailments, the results achieved are somewhat promising.

Apart from the global system evaluation, this year an extra evaluation has been introduced focused on measuring the impact of each system module in the final accuracy (called *the ablation tests*). In our specific case, we prepared five system configurations regarding this issue, Table 3 depicts explanations about the ablation tests carried out. All ablation test results are shown in Table 4.

Table 3: Ablation tests descriptions.

Test	Description
$ablation\_test_1$	just considering the system features derived from string-based similarities (see section 2.1)
$ablation\_test_2$	just considering the features from lexical-semantic analyses (see section 2.2)
$ablation\_test_3$	all system features except the WordNet-based ones (section 2.2.1)
$ablation\_test_4$	all features minus the ones derived from the frame semantic analysis (section 2.2.5)
$ablation\_test_5$	all features except the negation ones (section 2.2.2)

As a conclusion from the ablation tests, we can observe that the string-based inferences still have a strong influence in the entailment decision, and the lexical-semantic modules while reporting slight increase in accuracy, can establish their importance order as: "the FrameNet-based module", "the WordNet-based module" and "the Negation module", at least for the 2-way run. Regarding 3-way, the behaviour was a little bit different and was even confusing, we blame this on the lack of the system capabilities to distinguish between "UNKNOWN" and "CONTRADICTION" entailments. Therefore, subsequent work will be directed towards adjusting our inferences in managing the differences between these two kinds of entailments.

Table 4: Results obtained for the ablation tests over the 2-way and 3-way runs.

Run	**2-ways task**		**3-ways task**	
	Dev. corpus	Test corpus	Dev. corpus	Test corpus
$ALL_{inferences}$	0.6533	0.6283	0.6083	0.6
$ablation\_test_1$	0.635 (↓)	0.6183 (↓)	0.5983 (↓)	0.6033 (↑)
$ablation\_test_2$	0.6016 (↓)	0.595 (↓)	0.5583 (↓)	0.5683 (↓)
$ablation\_test_3$	0.6466 (↓)	0.62 (↓)	0.5983 (↓)	0.6033 (↑)
$ablation\_test_4$	0.635 (↓)	0.6167 (↓)	0.59 (↓)	0.6017 (↑)
$ablation\_test_5$	0.625 (↓)	0.6217 (↓)	0.5916 (↓)	0.605 (↑)

4 Conclusions

Throughout this paper, we have presented a system that recognizes entailment relations by merging shallow knowledge with more sophisticated knowledge derived from lexical-semantic inferences. Specifically, we use WordNet, detection of negation terms, NE recognition, verbs implications and frame semantic analysis to carry out the recognition task.

The system was applied over both the 2- and 3-way classification tasks, and although it was not designed to deal with three different sorts of entailments, the promising results obtained in this task encourage us to go on with this line and set the inferences for the 3-way classification problem.

The ablation-test results point out that it is not worth the effort made to find out ways that support the entailment decision using semantic resources. However, especially with FrameNet, we realized from our empirical studies that the robustness offered by semantic frames can potentially help in recognizing entailments, but unfortunately that is not very reflected in the results. Obviously, that is due to the limited coverage of FrameNet (although growing daily), the way we take advantage of the Frame-based knowledge, or both things. Therefore, our priority future work is to analyze why the richer knowledge provided by semantics, which is supposed to support the final decision, does not assist us very much, and how to overcome our ignorance about obtaining a proper linguistic modelling of the entailment phenomenon.

Acknowledgements

This research has been partially funded by the QALL-ME consortium, which is a 6^{th} Framework Research Programme of the European Union (EU), contract number FP6-IST-033860, by Generalitat Valenciana GV PROMETEO/2009/119 and by the Alicante University post-doctoral fellowship program funded by Fundación CajaMurcia (Spain).

References

[1] Eugene Agichtein, Walt Askew, and Yandong Liu. Combining semantic and lexical evidence for recognizing textual entailment (RTE4) task. In *Notebook Papers of the Text Analysis Conference, TAC 2008 Workshop*, Gaithersburg, Maryland, USA, November 2008. National Institute of Standards and Technology.

[2] Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluis Padró, and Muntsa Padró. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA*, Genoa, Italy, May 2006.

[3] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 86–90, San Francisco, California, 1998. Morgan Kaufmann Publishers.

[4] Alexandra Balahur, Elena Lloret, Manuel Palomar Óscar Ferrández, Andrés Montoyo, and Rafael Muñoz. The DLSIUAES Team's Participation in the TAC 2008 Tracks. In *Notebook Papers of the Text Analysis Conference, TAC 2008 Workshop*, Gaithersburg, Maryland, USA, November 2008. National Institute of Standards and Technology.

[5] Julio Javier Castillo and Laura Alonso i Alemany. An approach using named entities for recognizing textual entailment. In *Notebook Papers of the Text Analysis Conference, TAC 2008 Workshop*, Gaithersburg, Maryland, USA, November 2008. National Institute of Standards and Technology.

[6] Timothy Chklovski and Patrick Pantel. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain, 2004.

[7] Ido Dagan and Oren Glickman. Probabilistic textual entailment: Generic appied modelling of language variability. In *Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France, 2004.

[8] Katrin Erk and Sebastian Pado. Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC 2006: the 5 th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

[9] Óscar Ferrández, Rafael Muñoz, and Manuel Palomar. Studying the influence of semantic constraints in ave. In *CLEF 2008, Lecture Notes in Computer Science, to appear*, Aarhus, Denmark, September 2008.

[10] Danilo Giampiccolo, H.T. Dang, Bernardo Magnini, Ido Dagan, and Bill Dolan. The fourth pascal recognizing textual entailment challenge. In *Proceedings of the TAC 2008 Workshop*, National Institute of Standards and Technology, Gaithersburg, Maryland, November 2008.

[11] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The Third PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[12] Matthew A. Jaro. Probabilistic linkage of large public health data file. *Statistics in Medicine*, 14:491–498, 1995.

[13] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of International Conference on Research in Computational Linguistics*, 1997.

[14] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending VerbNet with Novel Verb Classes. In *Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy, June 2006.

[15] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[16] 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.

[17] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[18] Saul Needleman and Christian Wunsch. A general method applicable to the search for similarities in amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[19] Ted Pedersen, Siddhart Patwardhan, , and Jason Michelizzi. WordNet::Similarity - Measuring the Relatedness of Concepts. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 38–41, Boston, Massachussets, United States of America, 2004.

[20] Giuseppe Pirrò and Nuno Seco. Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In *On the Move to Meaningful Internet Systems: OTM 2008, OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Monterrey, Mexico, November 9-14, 2008, Proceedings, Part II*, volume 5332 of *Lecture Notes in Computer Science*, pages 1271–1288. Springer, 2008.

[21] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995.

[22] Álvaro Rodrigo, Anselmo Peñas, and Felisa Verdejo. Towards an entity-based recognition of textual entailment. In *Notebook Papers of the Text Analysis Conference, TAC 2008 Workshop*, Gaithersburg, Maryland, USA, November 2008. National Institute of Standards and Technology.

[23] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[24] Karen Sparck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[25] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* 2nd Edition, Morgan Kaufmann, San Francisco, 2005.