Exploiting Lexical Measures and a Semantic LR to Tackle Textual Entailment in Italian

Óscar Ferrández[1], Antonio Toral[2], and Rafael Muñoz[1]

[1] Natural Language Processing and Information Systems Group
Department of Computing Languages and Systems, University of Alicante
{ofe, rafael}@dlsi.ua.es
[2] Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche, Pisa, Italy
antonio.toral@ilc.cnr.it

Abstract. This paper discusses the participation of the University of Alicante and the *Istituto di Linguistica Computazionale* in the textual entailment exercise at EVALITA 2009. We present a system based on our previous experiences on the RTE Challenges. The system uses a machine learning classifier fed by features derived from lexical distances, part-of-speech information and semantic knowledge from SIMPLE-CLIPS, an Italian Language Resource. Although it was our first attempt in recognising entailment relations in Italian and the system was not thought in principle to deal with them, the results achieved encourage us to carry on doing research on this area. We obtain 58% accuracy when applying only lexical features. By considering also semantic knowledge derived from a Language Resource, accuracy reaches 64%.

Keywords: Textual Entailment, EVALITA 2009.

1 Introduction

EVALITA 2009 is the second evaluation campaign of Natural Language Processing (NLP) tools for Italian, which follows the previous success of EVALITA 2007. In both editions, the main aim is to promote the development of NLP technologies for Italian by providing a shared framework to evaluate such systems. EVALITA'09 comprises different text- and speech-based tasks such as Lexical Substitution, Entity Recognition, Textual Entailment, Spoken Dialogue, etc. We have focused our participation on the textual entailment recognition task.

Within the textual entailment exercise, participant systems have to detect unidirectional meaning implications between pairs of short texts. In such relations the meaning of one snippet must entail the meaning of the other, should this not occur the entailment relation does not hold. The snippet that permits the meaning inference is traditionally called T (the text) and the other, whose meaning is deduced, is named H (the hypothesis), as defined in [1].

In our participation, we present a system that integrates several inferences derived from different knowledge sources. The proposed approach is based on our

past participations in the Recognising Textual Entailment (RTE) Challenges for English [2, 3]. Obviously, some system inferences had to be removed since they used English resources such as VerbNet[3], VerbOcean[4], Named Entity Recognizers, etc. Therefore, what we wanted to check with our participation was if adapting to Italian some of the inferences studied along our RTE experiences, we can build an Italian RTE system reaching satisfactory results.

Additionally, we also made use of the semantic knowledge provided by the SIMPLE-CLIPS computational lexicon [4], attempting to establish relations between the senses (semantic units) and the ontological nodes (semantic types) of this resource that correspond to the words appearing in the pair text-hypothesis.

The remainder of the paper is structured as follows: Section 2 presents a detailed description of the approach; the EVALITA'09 results as well as the experiments carried out are shown in Section 3; finally, Section 4 gives some discussion and conclusions about the work done with our participation.

2 System Description

As briefly commented on the introductory section, aimed at achieving an approach based on our previous experiences on the RTE Challenges, we built our system focused on lexical deductions, basic inferences supported by part-of-speech (PoS) information and semantic implications using the SIMPLE-CLIPS lexicon. All the developed inferences are responsible for extracting a set of features that will be passed to a machine learning algorithm. For this issue, we use the Weka Framework [5], and as will be explained in section 3, we will show results using two different algorithms: Support Vector Machine and KStar. All the system's inferences will be profoundly explained in the next subsections.

2.1 The Lexical-based Component

Lexical overlappings are characterized by their simplicity and accuracy. Such techniques obtain quite promising results and in many cases are the base of lots of textual entailment systems. Hence, we considered very attractive the idea of integrating into our system a module focused on such techniques.

Our lexical-based component carries out the computation of several lexical distances between the lemmata belonging to the pair T-H (without considering Italian stop-words). Such distances are based on word co-occurrences as well as the context where they appear achieving a similarity factor between the target texts. In order to obtain the lemmata and the PoS information, we used the Italian configuration of Freeling toolkit [6]. The similarity scores will serve as features for our machine learning algorithm. The set of lexical measures contains:

- **Simple matching**: binary word overlap between the T's and H's lemmata.
- **Levenshtein distance matching**: similar to the simple matching, but in this case computing the Levenshtein distance between the lemmata.

[3] http://verbs.colorado.edu/ mpalmer/projects/verbnet.html
[4] http://demo.patrickpantel.com/Content/verbocean/

- **Smith-Waterman algorithm**: this algorithm instead of looking at each sequence in its entirety compares segments of all possible lengths and chooses whichever maximize the similarity obtaining optimal local sequence alignments. Further details can be found in [7]. In our experiments we set empirically the values 0.3, -1 and 2 for a gap, copy and substitution respectively.
- **Matching of consecutive subsequences**: it assigns the highest relevance to the appearance of consecutive subsequences, considering those from length two until the length in words of the hypothesis. Therefore, a matching procedure is performed between the consecutive subsequences generated. Non-consecutive subsequences are not taken into account, the same relevance is assigned to all consecutive subsequences with the same length and, the longer the subsequence is, the more relevant it will be considered.
- **Jaro distance**: it comes from the work presented in [8] and measures the similarity between two strings taking into account spelling derivations:

$$d_j(s_1, s_2) = \frac{m}{3 \cdot |s_1|} + \frac{m}{3 \cdot |s_2|} + \frac{m - t}{3 \cdot m} \tag{1}$$

  being s_1 and s_2 the strings to be compared, $|s_1|$ and $|s_2|$ their respective lengths, m the number of matching characters considering only those are not farther than $[\frac{\max(|s_1|,|s_2|)}{2}] - 1$ and t the number of transpositions computed as the number of matching (but different) characters divided by two.
- **Cosine similarity**: is a common vector-based similarity. The input strings are transformed into vector space and it is computed as follows:

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||} \tag{2}$$

- **Rouge measures**: ROUGE measures have already been tested for automatic evaluation of summaries and machine translation [9]. For this reason, and considering the impact of n-gram overlap metrics in textual entailment, we consider very interesting the idea of implementing these measures. Specifically, we implemented the ROUGE-N measure with n equal to 2 and 3, the ROUGE-L as an LCS-based F-measure, the ROUGE-W measure as an F-measure based on WLCS and the ROUGE-S measure as an F-measure based on skip-bigrams.[5]

In addition to these lexical distances, we also wanted to take into account the knowledge supplied by the PoS information (grammar category, singular, plural, person, tense, etc.). After analysing the training corpus, we realized that in some cases T and H are somewhat similar and the entailment relation does not hold due to slight modifications regarding the PoS tags. Specifically, we notice false entailment relations because of different verb tenses between T and

[5] LCS means Longest Common Subsequence.
WLCS means Weighted LCS, which is a LCS modification that memorizes the length of consecutive matches encountered.
For those ROUGE measures that are F-based, the length of T was used for the recall, the length of H for the precision and the β parameter was set to one.

H. Therefore, two more similarity values were inferred: (1) a simple overlapping between the PoS tags; and (ii) a value showing the comparison degree between the verb tenses appearing in both T and H. Both values will be considered as system features.

2.2 The SIMPLE-CLIPS Language Resource

SIMPLE-CLIPS [4] is an Italian computational lexicon based on the Generative Lexicon theory. This resource is made up of different layers: phonologic, morphologic, syntactic and semantic. The semantic layer, the relevant one for the current research, contains about 55,000 word senses organised in an ontology made up of 153 nodes.

The ontology is made up of a hierarchy of nodes called semantic types. There are two types of nodes, simple types, which are identified by only a one-dimensional aspect of meaning (formal) expressed by hyperonymic relations, and unified types, for which additional dimensions of meaning (i.e. the three qualia elements agentive, constitutive and telic) are needed.

The lexicon consists of word senses (called semantic units) structured in terms of the semantic system type defined by the ontology. The entries contain different types of semantic information such as mandatory and optional features and relations, predicates including roles, restrictions, etc. To express the relationships holding between semantic units, the model provides 138 different semantic relations, out of which 60 regard extended qualia relations. These relations allow for the expression of very fine-grained distinctions both for structuring the information regarding the componential aspect of word meanings and to capture the nature of the relationships holding among word senses. The lexicon instantiates more than 80,000 relations.

We have used in our textual entailment system semantic information regarding the (i) semantic types and (ii) semantic units. Given two lemmas we obtain:

(i) the corresponding semantic units, and following their relations obtain the shortest path that connects them in the SIMPLE-CLIPS semantic graph[6].
(ii) the corresponding semantic types, and following the is-a relations obtain the shortest path that connects them in the formal taxonomy of the ontology.

Regarding (i), we sum the minimum distances between semantic units that appear in H with respect to the semantic units that appear in T. The sum is normalized by the total number of semantic units present H, but if a lemma in H has no corresponding semantic unit it is not taken into account for the normalization. For (ii), we sum the minimum distances between semantic types in H and T as in (i) and, additionally, we calculate the overlapping of semantic types in H and T (as done for PoS tags). As for the previous inferences, each of them is considered as a system feature.

[6] The vertices of the graph are the semantic units and the edges the relations between pairs of semantic units

3 EVALITA 2009 Results

Table 1 summarizes the results obtained with a 10-fold cross validation over the development data and the final system's accuracy using the test-blind corpus provided by the organizers.

Table 1. Results obtained for the EVALITA Textual Entailment 2009 track.

Run	**Development corpus**	**Test-blind corpus**
	10-fold cv. accuracy	**accuracy**
KStar_Lexical_Features	0.695	0.56
KStar_Lexical&Semantic_Features	0.7375	**0.64**
SVM_Lexical_Features	0.6725	0.58
SVM_Lexical&Semantic_Features	0.675	0.57

We carried out four experiments, which were split up into two groups depending on the machine learning algorithm used (i.e. $KStar\_xxx$ vs. SVM_xxx). Each group contains two different experiments: (i) just considering the features from the lexical distances; (ii) all lexical-driven features plus the ones derived from the PoS information as well as those obtained from SIMPLE-CLIPS. With these experiments we wanted to check the impact in taking the entailment decision of the lexical module, the PoS information and the semantic lexicon-based inferences.

4 Discussion and Conclusions

The results achieved are in the line of the most state-of-the-art textual entailment systems for English (see [10]), and they also point out that even though the system was not designed to deal with entailment pairs in Italian, its behaviour was somewhat promising. Moreover, taking into account that the inferences implemented are quite simple, we are even more satisfied with our participation, and this fact encourages us to go on to further research in this area.

In the training phase, apart from assessing the importance of each system feature (lexical, PoS-driven, semantic), we also wanted to check how our set of features works with different machine learning algorithms. So, we made some experiments using SVM, decision trees, Bayesian networks and instance-based and rule-based learners. These experiments revealed that the subset of lexical features worked better when the classifier was SVM. However, when this set was enriched with the rest of inferences, it was the KStar instance-based learner which reported the best performance.

In a nutshell, this paper has presented a system that recognizes entailment relations by merging shallow lexical knowledge with more sophisticated knowledge derived from semantic inferences. We obtain 58% accuracy when applying

only lexical features. By considering also semantic knowledge derived from a Language Resource, accuracy reaches 64%. We can conclude that the lexical inferences have a huge influence in the entailment decision, while the semantic ones report a slight increase in accuracy (around 10%). However, we strongly believe that the robustness offered by semantics is the proper way to solve entailments, and the positive results obtained by applying into the system knowledge related to ontology nodes and semantic relations point in this direction. Therefore, our priority future work is to analyze how to integrate richer knowledge into the system, in order to obtain a suitable modelling of entailment relations.

Acknowledgments. This research has been funded by the EU Commission under the projects QALL-ME (FP6-IST-033860) and KYOTO (ICT-2007-211423), and by the Alicante University post-doctoral fellowship program funded by Fundación CajaMurcia.

References

1. Glickman, O.: Applied Textual Entailment. PhD thesis, Bar Ilan University (2006)
2. Ferrández, Ó., Micol, D., Muñoz, R., Palomar, M.: A Perspective-Based Approach for Solving Textual Entailment Recognition. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 66–71. Association for Computational Linguistics (2007) 66–71
3. Balahur, A., Lloret, E., Óscar Ferrández, Andrés Montoyo, M.P., Muñoz, R.: The DLSIUAES Team's Participation in the TAC 2008 Tracks. In: Notebook Papers of the Text Analysis Conference, TAC 2008 Workshop. Gaithersburg, Maryland, USA (2008)
4. Ruimy, N., Monachini, M., Distante, R., Guazzini, E., Molino, S., Ulivieri, M., Calzolari, N., Zampolli, A.: Clips, a multi-level italian computational lexicon: A glimpse to data. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). Las Palmas de Gran Canaria, Spain (2002)
5. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco (2005)
6. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M.: 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). Genoa, Italy (2006)
7. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology, vol. 147, pp. 195–197 (1981)
8. Jaro, M.A.: Probabilistic linkage of large public health data file. Statistics in Medicine, vol. 14, pp. 491–498 (1995)
9. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out: Proceedings of the Association for Computational Linguistics Workshop, pp. 74–81. Barcelona, Spain (2004)
10. Giampiccolo, D., Dang, H., Magnini, B., Dagan, I., Dolan, B.: The fourth pascal recognizing textual entailment challenge. In: Proceedings of the TAC 2008 Workshop, National Institute of Standards and Technology. Gaithersburg, Maryland (2008)