Evaluation of a Semantically Oriented Dependency Grammar for Italian at EVALITA 2009

Marcella Testa, Andrea Bolioli, Luca Dini, and Giampaolo Mazzini

CELI, via San Quintino 31,
10131 Torino, Italy
{testa,abolioli,dini,mazzini}@celi.it

Abstract. This paper describes CELI's participation to the EVALITA 2009 dependency evaluation task, which was based on a symbolic dependency grammar for Italian. The discussion on the results achieved by such a grammar emphasizes the fact that they are obtained by linearizing the CoNLL input and parsing it from scratch, thus generating extra errors due to lexicon look-up and POS tagging. In the final section we provide some explanation of the fact that the scores obtained in the pilot subtask were lower than the ones for the main subtask. Overall, however, the results of the participation prove that a rule approach is viable and sustainable for real life applications.

Keywords: Dependency Parsing, Italian Syntax, Grammar Evaluation.

1 Introduction

In this paper we describe CELI's participation in the Evalita 2009 experiment. CELI participated in the syntactic task, in particular in the dependencies track, using both the TUT (Turin University Treebank) for the main subtask, and the TANL (Text Analytics and Natural Language) for the pilot subtask.

CELI's interest in the participation was mainly to evaluate the performance of an Italian grammar in a real context. As we will show, the assumption of the application of the grammar in a real environment caused some decrease in scoring with respect to the average of the participants. In the last section of this paper we will discuss, in detail, the differences in scores that were obtained between the TUT and the TANL corpora.

2 The Experiment

The system that was tested at Evalita is based on three main components. The grammatical core is a grammar-based dependency parser that will be described in the following section. The second is a component in charge of translating the parser's internal representation into CoNLL data format. The third component is the one that takes as input the CoNLL structure representing our parser output, and maps it to the specific structure of TUT and TANL.

2.1 The Grammar

The grammar we used was encoded by using XIP, the Xerox Incremental Parser [1], and was developed in a period of about six months of a full time equivalent. The grammar for Italian is based mainly on assumptions drawn by the dependency grammar [2]. However, it should be noticed that in general the interpretations of such assumptions are more semantic than syntactic in nature. For instance, almost all dependencies, contrary to standard dependency grammars, hold between semantic heads, rather than syntactic heads, as it is usually assumed. As we will see, our interpretation added some complexity to our participation in Evalita.[1]

2.2 The Results

The following table shows the results that have been achieved by CELI, as computed by eval.pl script:

TASK	LAS	UAS	LA[2]
TUT	68.00	72.97	77.95
TANL	57.81	64.10	73.86

If we compare them to the average of the other participants, we can see that these results are below the average scores in both TUT and TANL experiments. The average of the other participants (without taking into account our results) are presented in the following table:

TASK	LAS
TUT	85.86
TANL	78.96

In the next two sections we will first explain the differences with respect to the TUT corpus, and then we will try to give an explanation of the different scores achieved comparing the TUT and the TANL experiments.

[1] In general, concerning grammar development we would like to express our gratitude to the support provided by the Parsing and Semantics Group of the Xerox Research Centre Europe. We also thank Sigrid Maurel, who carried on the development of the first version of the Italian dependency grammar.

[2] Label accuracy score.

3 Explanation (TUT)

In general, from an application-oriented point of view, we estimate that the results we achieved are quite satisfactory, as they are based on a "real-life" assumption, as we will see.

However if we compare our LAS score (68.00%) with the ones of the other participants, we can see that it is in general below the average (all participants achieved results higher than 80%).

Now there could be two different explanations of this phenomenon. We could, of course, formulate the hypothesis that, in general, statistical parsers perform better than symbolic ones. However this claim is disconfirmed by the fact that in 2007 the dependency parser of Turin University [3] obtained results that are higher that those we obtained (with a LAS of 86.94%). Therefore some other reasons need to be found.[3]

In general, we believed that the major cause of the differences between our parser and the average results is due to the fact that, rather than considering the partially annotated text (i.e. containing part-of-speech, syntactic features and morphological information) as the input for our parser, we basically linearized the input text and we performed analysis from scratch starting from tokenization up to dependency parsing. The major consequence of this process is that we lose all information concerning part of speech disambiguation, thus adding internal disambiguation errors to pure dependency parsing errors. The second point is that we perform our own access to the lexicon, thus introducing errors which are due to lexical gaps or anyway bad encoding of lexical items.

On top of that, as we said, our grammar is based on an assumption which is completely different from the one of the TUT, that is the one of computing dependencies among semantic heads. This fact forced us to write the special mapping component. It is clear that in this programmatic mapping of our semantic structures to TUT syntactic structure, some errors might have occurred.

Finally, our parser has been built with the assumption of not performing attachment choices. For instance, a PP can be tied both to nominal and verbal node; this assumption is justified by the fact that the PP-attachments can be often decided in a practical application context, according to the domain. However, in the context of an evaluation, this ambiguity is not accepted, and we were in many cases forced to perform a random choice that decreased the precision.

[3] It is true that, as the organizers claims "it has been developed in parallel with the TUT and so we can guess a certain influence over the annotators of the gold standard of the test set". However we believe that such an influence does not invalidate the brilliant performance of such a parser. In the same evaluation [4] reports 47.62 LAS for a rule-based approach.

4 Differences between TUT and TANL

By considering the results that have been achieved by all the participants, we note that, in general, all performances decreased in the pilot task with respect to the main task.

However, in our case, the difference between the two tasks is higher than the average: while all the other participants decrease their score by 6.9%, our LAS on TALN decreases by 10.19%. It is interesting to investigate the cause for such decrease. In this respect, we identified three possible reasons, namely biased development, different structuring of information and errors in mapping.

Concerning biased development, it is important to consider that the XIP Italian grammar is based on a cycle of implementation, verification and debugging which used a version of the TUT as a golden standard. Therefore it might be the case that, to a certain extent, the grammar is biased towards such a corpus. This bias can be classified into two different categories:

- the first one is a real bias, in the sense that certain phenomena *could* have been modeled according to the frequencies they occurred in the TUT;
- the second one is an issue of phenomenon mapping, in the sense that certain phenomena, such as coordination, have been *really* modeled according to the TUT assumptions. These instances naturally mapped better in an evaluation experiment based on such a corpus. For instance, the average of the scores for coordinative dependencies in TUT shows a precision of 79.3% and a recall of 70%; the same average for TANL is, respectively, only 61.6% and 57.3%.

Concerning the "structural" difference, the difference in score between TUT and TANL might be due to a finer grained information contained in the TANL corpus with respect to the TUT one.[4] Most notably, this difference is represented by the distinction of complements and adjuncts into locative, temporal and all the other kinds of modifiers.

Finally, while the mapping from our grammar to the TUT has been continuously improved over a six-month development period, the mapping of our dependencies to the TANL linguistic assumptions was a last week endeavor, and a greater investment of time in the latter would have probably yielded more favorable results.

5 Conclusions

In this paper we described CELI's approach to the EVALITA 2009 task on dependency parsing. We gave details about our approach, which is completely based on human encoding of syntactic grammars. The results of the test prove that such an approach is viable and sustainable for real life applications.

[4] This claim is not to be taken in an absolute sense, but relatively to our grammar: certain distinctions encoded by TALN were absent from our Italian Grammar, which presented a granularity more similar to the one of TUT corpus in CoNLL format.

References

1. Aït-Mokhtar, S., Chanod, J.P., Roux, C.: Robustness below shallowness: Incremental deep parsing. Special Issue of the Natural Language Engineering Journal on Robust Methods in Analysis of Natural Language Data, vol. 8, pp. 121--144 (2002)
2. Nivre, J.: Dependency Grammar and Dependency Parser. MSI report 05133. School of Mathematics & Systems Engineering, Växjö University, http://w3.msi.vxu.se/~nivre/papers/05133.pdf (2005)
3. Lesmo, L.: The rule-based parser of the NPL group of the university of Torino. Intelligenza artificiale, vol. 4, pp. 46--47 (2007)
4. Zanzotto, F.M.: Lost in grammar translation. Intelligenza artificiale, vol.4, pp. 42--43 (2007)