Comparing Italian parsers on a common treebank: the Evalita experience

C. Bosco*, A. Mazzei*, V. Lombardo*, G. Attardi+, A. Corazza%, A. Lavelli#, L. Lesmo*, G. Satta*, M. Simi*

* Università di Torino, # FBK-irst - Trento, % Università Federico II - Napoli, + Università di Padova, # Università di Pisa
Address1, Address2, Address3
{bosco,mazzei,vincenzo,lesmo}@di.unito.it, {attardi,corazza}@unipi.it, {lavelli,fbk,eu,lesmo,na.infn.it, satta}@dei.unipd.it

Abstract

The Evalita ‘07 Parsing Task has been the first contest among parsing systems for Italian. It is the first attempt to compare the approaches and results of the existing parsing systems specific for this language using a common treebank annotated both in dependency and constituency-based format. The paper describes the parsing competition and proposes an analysis of the results, which takes into account other experiences in treebank-driven parsing for Italian and other Romance languages. It focuses on datasets, parsing paradigms and approaches applied on Italian French and German.

1. Introduction

By providing a very large set of syntactically annotated sentences, the Penn Treebank has played an invaluable role in enabling the development of state-of-the-art parsing systems (Ratnaparki, 1997; Charniak, 1997; Collins, 1999). But the strong focalization on Penn Treebank, and more specifically on the Wall Street Journal portion of this treebank, has left open several questions on parsers portability.

The application of parsing methods to different languages and treebanks is currently considered as a crucial and challenging task, and system porting across text genres, languages and annotation formats should be a research problem in itself. The validation of existing parsing models, in fact, strongly depends on the possibility of generalizing their results on corpora other than those on which they have been trained and tested.

For constituency-based parsing, strong empirical evidence demonstrates that results obtained on a particular treebank are unportable on other corpora. For instance, (Gildea, 2001) shows that the results obtained on the Wall Street Journal section of the Penn Treebank are not reproducible on the Brown Corpus, which is annotated according to the same format but contains texts featured by different genre. Other works showed the irreproducibility across English and other languages even if annotated in Penn-like formats, see e.g. (Collins et al., 1999) on Czech, (Dubey and Keller, 2003) on German, (Levy and Manning, 2003) on Chinese, (Corazza et al., 2004) on Italian. While, e.g. (Kübler, 2005; Kübler and Prokič, 2006; Maier, 2006) on Negra and TubaDZ treebanks, show that parsing results vary according to the features of the annotation schema applied to the same corpus of sentences, i.e. dependency or constituency-based.

For dependency parsing, the results of the 2006 and 2007 CoNLL multilingual shared task (Buchholz and Marsi, 2006; Nivre et al., 2007a) together with those reported in (Nivre et al., 2007b; Chanev, 2005), showed that it is as robust as the constituency parsing, but equally affected by the problem of irreproducibility of results across corpora and languages.

The aim of the EVALITA ’07 Parsing Task (EPT), held in Frascati (Rome) in September 2007 (Bosco et al., 2007), was to assess the current state-of-the-art in parsing Italian by encouraging the application of existing parsing models to this language, and to contribute to the investigation on the causes of this irreproducibility. It allowed to focus on Italian by exploring both different paradigms, i.e. constituency and dependency, and different approaches, i.e. rule-based and statistical. In fact, the task was composed of subtasks with separate development datasets and evaluations for constituency and dependency. Therefore, the EPT can be seen as the first picture of the problems that lie ahead for Italian parsing and the kind of work necessary for adapting existing parsing models to this language.

The paper presents an analysis of the results that goes beyond the limits of the event. The next section presents the development and test datasets, and the evaluation metrics applied in EPT. The other two sections respectively propose the results obtained by participant parsing systems and an analysis of these results also by developing comparisons with parsing experiences on other languages and in similar contests.

2. Task, datasets and evaluation metrics

The parsing task for EPT is defined as the activity of assigning a syntactic structure to a given Italian Part of Speech (PoS) tagged sentence using a fully automatic parser. The syntactic structure has to be described according to one of two annotation schemes presented in the development set, one for dependency and one for constituency subtask.

The annotation schemes, data sets and standard evaluation metrics applied in EPT are described in the rest of this section.

2.1. Development and test datasets

The reference treebank for EPT is the Turin University Treebank (TUT), which is available both in dependency
and constituency format (see at the TUT web site for a free download http://www.di.unito.it/~tutreeb).

The native annotation schema of the treebank is dependency-based. It follows the major tenets of Hudson’s dependency grammar (Hudson, 1984), but includes null elements for the representation of particular phenomena, such as non-projective structures and pro-drops. Moreover, the treebank features a rich set of grammatical relations (i.e. around 250 relations) developed according to the Augmented Relational Structure (Bosco and Lombardo, 2004). Each of these relations can, in fact, include three different components, i.e. morpho-syntactic, functional-syntactic and syntactic-semantic. For instance in the relation PREP-RLMOD-TIME, PREP the morpho-syntactic component, RMOD the functional-syntactic component, and TIME the syntactic-semantic one. This allows for a representation which is scalable at different degrees of specificity. For instance, by selecting only the functional-syntactic component of each relation, we can reduce the cardinality of the relation set from 250 (fully-specified) to 74 (specified only from the functional-syntactic point of view) items.

In order to increase the comparability with other works and to make the data more adequate for the application of standard measures for the evaluation of parsing results, some efforts have been devoted by organizers to the standardization of data. In particular, the development set for EPT has been made available:

- in the above mentioned annotation with a reduced relation set,
- in a null elements free annotation
- in the 10-column standard CoNLL format.

By applying automatic procedures to the native annotation, as described in (Bosco and Lombardo, 2006), the dependency treebank has been converted in some constituency-based formats, among which a Penn-like one, called TUT-Penn. The TUT-Penn format includes, as usual for constituency-based annotations, null elements. Nevertheless, even if it implements the same structure of Penn data, it uses a specific PoS tag set. In fact, as in other cases of treebank conversion (Collins et al., 1999), the use of a specific set of PoS tags, which are derived by reduction from the TUT original PoS tags, has been preferred to the original Penn PoS tags since they better represent the inflectional richness of Italian.

The full TUT corpus, available in all the above described annotations and formats, has been offered to the EPT participants as development corpus. It currently consists in 2,000 sentences that correspond to about 58,000 annotated tokens. In order to allow for comparison of results across text genres, the treebank is organized in two subcorpora of one thousand sentences each, i.e. the Italian legal Code (47.5% of tokens) and Italian newspapers (52.5% of tokens).

The test set consists instead of 200 new sentences, i.e. 100 from newspapers and 100 from Italian legal Code, in order to represent a text genre balancement similar to that in the development set, and thus to allow for separate evaluations on the different genres.

2.2. Evaluation metrics
The evaluation of dependency results is based on the three CoNLL standard metrics:

- Labeled Attachment Score (LAS), the percentage of tokens with correct head and relation label;
- Unlabeled Attachment Score (UAS), the percentage of tokens with correct head;
- Label Accuracy (LAS2), the percentage of tokens with correct relation label.

For constituency, the evaluation is based on standard PARSEVAL measures:

- Brackets Precision (Br-P), the percentage of found brackets which are correct;
- Brackets Recall (Br-R), the percentage of brackets correct which are found;
- Brackets F (Br-F), the composition of the previous two measures that can be calculated by the following formula: $2 \times (P \times R)/(P + R)$.

3. Participants and results
In this section, we describe the systems that particiapte to EPT and their results.

3.1. Submissions and results
Test runs were submitted to EPT by 8 participants\(^1\), among which 5 from Italy and the other from foreign countries (India, Germany, USA), 7 from academic institutes. The submissions of 6 teams is for dependency parsing, and the remaining 2 for constituency parsing. Nobody for both subtasks. In the result tables, one for dependency and one for constituency, systems are identified by the institution name and by the last name of the first team member separated by underscore, as in (Bosco et al., 2007).

3.1.1. Dependency subtask
For the dependency subtask, the participant parsers are the following.

The parser UniTo_Lesmo includes chunking followed by attachment of verb dependents driven by both rules manually developed and data about verbal subcategorization. It is a rule-based parser developed in parallel with the TUT and tuned on the dataset. The UniPi_Attardi (of the team composed by Attardi and Simi), called DeSR, is a multilingual deterministic shift-reduce dependency parser that handles non-projective dependencies incrementally and learns by a second-order

\(^1\)Among participants, five are single authors, while the others are teams.
The IIIT_Mannem is an online large margin based training framework for deterministic parsing using Nivre’s shift-reduce parsing algorithm.

The UniStuttIMS_Schiehlen uses Eisner’s bottom-up chart-parsing algorithm for inference and online passive aggressive algorithms for learning; it produces non-projective labelled trees.

The UPenn_Champollion system (by the team composed by Champollion and Robaldo) is a bidirectional dependency parser which does a greedy search over the sentence and picks the relation between two words with the best score each time and builds the partial tree.

The FBKirst_Pianta is instead a left corner parser for Italian, based on explicit rules manually coded in a unification formalism.

The details of their results are described in table 2.

Table 1 describes the details of their results according to the above defined standard measures and shows that the best scores for this task have been obtained by the UniTo_Lesmo.

Table 1: Dependency parsing subtask evaluation

<table>
<thead>
<tr>
<th>LAS</th>
<th>UAS</th>
<th>LAS2</th>
<th>Participant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>86.94</td>
<td>90.90</td>
<td>91.59</td>
<td>UniTo_Lesmo</td>
<td>1-1-1</td>
</tr>
<tr>
<td>77.88</td>
<td>88.43</td>
<td>83.00</td>
<td>UniPi_Attardi</td>
<td>2-2-2</td>
</tr>
<tr>
<td>75.12</td>
<td>85.81</td>
<td>82.05</td>
<td>IIIT_Mannem</td>
<td>3-4-3</td>
</tr>
<tr>
<td>74.85</td>
<td>85.88</td>
<td>81.59</td>
<td>UniStuttIMS_Schiehlen</td>
<td>4-3-4</td>
</tr>
<tr>
<td>85.46</td>
<td>*</td>
<td>*</td>
<td>UPenn_Champollion</td>
<td><em>-</em></td>
</tr>
<tr>
<td>47.62</td>
<td>62.11</td>
<td>54.90</td>
<td>UniRoma2_Zanzotto</td>
<td>5-6-5</td>
</tr>
</tbody>
</table>

Table 2: Constituency parsing subtask evaluation

<table>
<thead>
<tr>
<th>Br-R</th>
<th>Br-P</th>
<th>Br-F</th>
<th>Errors</th>
<th>Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>70.81</td>
<td>65.36</td>
<td>67.97</td>
<td>26</td>
<td>UniNa_Corazza</td>
</tr>
<tr>
<td>38.92</td>
<td>45.49</td>
<td>41.94</td>
<td>48</td>
<td>FBKirst_Pianta</td>
</tr>
</tbody>
</table>

Table 2: Constituency parsing subtask evaluation

The last part of this section focuses instead on the parsing approaches applied in EPT regardless of the subtask.

4. Analysis and discussion of results

In this section, the results obtained in EPT will be compared with those obtained for Italian by other data-driven parsing systems applied on it. We will present an analysis for each subtask and in the comparison, we will focus on the effects on parsing results of various parameters, but, in particular, of differences in the dataset size and annotation. Therefore, among the scores for Italian, we will take into account those based both on TUT and on other existing treebanks for the same language, namely the Italian Syntactic Semantic Treebank (ISST) (Barsotti et al., 2001). ISST is a treebank that implements an annotation schema different from that of TUT, with a syntactic annotation distributed over two levels, the constituent structure and the functional relations level where 22 dependency relations are attested.

The best scores for this task have been obtained by the UniTo_Lesmo.

First of all, the results of the EPT dependency subtask (see at table 1) may be compared with those of the MaltParser on Italian language. The MaltParser, described in (Nivre et al., 2007b), implements a language-independent data-driven approach which has been successfully applied to (LISTA LINGUE e DIRE SU QUALI MALI HA I RISULTATI BEST). As in EPT, the experiments performed for Italian by the MaltParser are based on TUT annotated according to the pure dependency annotation schema and with the reduced relation set (including 74 relations). Nevertheless, the data in the experiments with MaltParser are 1,500 sentences (that correspond to around 42,000 annotated tokens), rather than 2,000 sentences like in EPT (i.e. 58,000 annotated tokens). It is not surprising that the reduced size of the dataset results, in the MaltParser experiments, in lower scores than in EPT, i.e. 75.7 (versus 86.94 in EPT) for LAS and 82.9 (versus 90.90 in EPT) for UAS.

Since Italian was among the languages taken into account for the CoNLL07 multilingual dependency parsing shared task (Buchholz and Marsi, 2006), our comparison can be also extended to the results presented at the last edition of this competition among parsing systems. Italian was considered among the languages parsed with higher scores, i.e. achieving LAS between 84.40 and 89.61, together with Catalan, Chinese and English. In CoNLL, the dataset for Italian included a larger amount of sentences than in EPT, namely 3,100 which correspond to around 71,000 annotated tokens rather than 2,000 (i.e. 58,000 annotated tokens). But, by contrast with EPT, here the data come not from TUT but from ISST. This allows for a comparative analysis of the effect of annotation on parsing results. The best scores for Italian were 84.40 (versus 86.94 in EPT) for LAS, obtained by the parser described in (Hall et al., 2007), and 87.91 (versus 90.90 in EPT) for UAS achieved by the parser described in (Nakagawa, 2007). These results are still lower, but closer to those presented in EPT, and, moreover, they have been both obtained by the application of systems that implement the combination of more parsers. A more adequate comparison should therefore refer to the best performing single parser system, that in CoNLL was that by
Titov and Henderson (Titov and Henderson, 2007), which achieved yet lower scores, i.e. 82.26 (versus 86.94 in EPT) for the LAS score and 86.26 (versus 90.90 in EPT) for UAS. From a theoretical point of view, the difference of results in CoNLL shared task and in EPT can be motivated both by the application of different parsing models and by the different data set. In practice, the availability of results for parsers that participated to CoNLL and EPT both, can be of some help in find more precise motivations for these differences in results. Table 3 shows the results achieved by UniPli_Attardi, IIT_Mannem and UniStuttIMS_Schiehlen, in both CoNLL and EPT.

All the three systems obtained higher scores for LAS in CoNLL than in EPT, i.e. 81.34 versus 77.88 for UniPli_Attardi, 78.67 versus 75.12 for IIT_Mannem, and 80.46 versus 74.54 for UniStuttIMS_Schiehlen. This can be interpreted as a confirmation of the trivial fact that the performance of parsing systems are influenced by the features of the reference treebank. In particular, the increased amount of relations in TUT in EPT with respect to ISST (74 versus 31 relations), together with reduced size of the dataset (42,000 versus 71,000 tokens), can motivate the difference in scores. Nevertheless, all the three systems obtained also an higher UAS in EPT than in CoNLL. This can be interpreted as a further confirmation of the fact that the pure dependency annotation schemes, like that of TUT, can be considered as more adequate for the representation of Italian language, following the interpretation proposed also in (Chaney, 2005) for experiments on TUT.

A further evidence for this interpretation of results derives from the experiments described in (Attardi and Simi, 2007) where the same parser DeSR applied in the EPT, i.e. UniPli_Attardi, is used in the same EPT task by exploiting a smaller set of (less specialized) TUT relations (31 versus 74 relations), and a better result is achieved for the LAS score (i.e. 83.27) than in the official EPT task (i.e. 77.88).

Recent experiments using the IDP parser developed by Henderson-Titov, previously cited, result in further evidence for our interpretation. This system achieved for LAS 76.79 on TUT dataset and 82.26 on ISST; while for UAS 88.13 on TUT and 86.26 on ISST.

The results obtained for dependency parsing at the EPT can be therefore considered as satisfactory, since they are higher or very closer to the state-of-the-art. Moreover, they offered a valuable experimental evidence to previously formulated hypothesis, namely the adequateness of dependency parsing approaches to Italian. This is in line with similar hypothesis formulated for other languages featured by free word order, like German or Czech.

4.2. Constituency subtask

For the constituency parsing subtask, the results of EPT are less meaningful than those for dependency, because only two systems participate to this subtask and because there are less data available on this kind of parsing applied to Italian language. The comparison will therefore mainly refer to English, which remains, together with other fixed word order languages (like Chinese), the reference language for constituency-based parsing approaches.

In (Corazza et al., 2004) the same parser used in EPT (i.e. UniNa_Corazza) was run on a dataset composed by about 3,000 sentences from the ISST, achieving scores definitely lower than on a subset of the WSJ of comparable size (see at table 4). The worse results on Italian with respect to English are confirmed in EPT, and the smaller dataset in EPT results in lower scores than on ISST (Corazza et al., 2007). However, further experiments after normalizing multiword expressions (cause of the errors reported in table 2) produced better results, also higher than on ISST (71.73 for R, 69.88 for P and 70.79 for F) (Corazza et al., 2007). We expect better results on the new release of the TUT-Penn currently under development.

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPT (Italian)</td>
<td>70.81</td>
<td>65.36</td>
<td>67.97</td>
</tr>
<tr>
<td>ISST (Italian)</td>
<td>68.40</td>
<td>68.58</td>
<td>68.49</td>
</tr>
<tr>
<td>WSJ (English)</td>
<td>84.02</td>
<td>83.41</td>
<td>83.71</td>
</tr>
</tbody>
</table>

Table 4: Constituency parsing subtask evaluation

4.3. Parsing approaches

DA FARE

About the parsing approach, the systems presented in both EPT subtasks are all statistics-based except three that are rule-based, i.e. UniTo_Lesmo, FBK_Schirrip and UniRoma2_Zanzotto. Statistics-based parsers have achieved notable results although the development set is smaller than in CoNLL07, while the different tuning of the UniRoma2_Zanzotto rule-based parser can possibly explain the relatively poor performance with respect the UniTo_Lesmo.

5. Conclusions and future works

DA FARE

Even if different standard measures and different number of participants to the dependency and constituency subtasks make a direct comparison difficult, the EPT results contribute to the investigation on parsing system portability by showing a higher distance from the state-of-the-art for constituency than for dependency parsing for Italian. Moreover, the availability of a Penn-like annotation for Italian in EPT has allowed for a comparison between parsing performance on different Italian treebanks and confirmed that dependency structures seem to be more adequate for this language. An important consequence of the Evalita activity has been to strengthen the interactions among groups working on Italian parsing. Hopefully, this will lead to a common effort towards an in-depth comparison of annotation schemes and to the development of larger integrated resources.

6. References

Table 3: Comparison between EPT and CoNLL07 LAS and UAS, ordered according to the EPT LAS scores.

<table>
<thead>
<tr>
<th>Participant</th>
<th>LAS</th>
<th>UAS</th>
<th>CoNLL EPT</th>
<th>CoNLL EPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniP_{\text{Attardi}}</td>
<td>81.34</td>
<td>77.88</td>
<td>85.54</td>
<td>88.43</td>
</tr>
<tr>
<td>IIIT_{\text{Mannem}}</td>
<td>78.67</td>
<td>75.12</td>
<td>82.91</td>
<td>85.81</td>
</tr>
<tr>
<td>UniStuttIMS_{\text{Schiehlen}}</td>
<td>80.46</td>
<td>74.85</td>
<td>84.54</td>
<td>85.88</td>
</tr>
</tbody>
</table>


