Evalita'09 Parsing Task: constituency parsers and the Penn format for Italian

Cristina Bosco, Alessandro Mazzei, and Vincenzo Lombardo

Dipartimento di Informatica, Università di Torino, Corso Svizzera 185, 10149, Torino
{bosco,mazzei,vincenzo}@di.unito.it

Abstract. The aim of Evalita Parsing Task is at defining and extending the state of the art for parsing Italian by encouraging the application of existing models and approaches. Therefore, as in the first edition, the Task includes two tracks, i.e. dependency and constituency. This second track is based on a development set in a format, which is an adaptation for Italian of the Penn Treebank format, and has been applied by conversion to an existing dependency Italian treebank.
The paper describes the constituency track and the data for development and testing of the participant systems. Moreover it presents and discusses the results, which positively compare with those obtained for constituency parsing in the Evalita'07 Parsing Task.

Keywords: Parsing evaluation, constituency, Italian, treebank, Penn Treebank format

1 Introduction: motivations of the constituency Parsing Task

The general aim of the Evalita Parsing Task contests is at defining and extending the current state of the art in parsing Italian with reference to the existing resources, by encouraging the application of existing models to this language.

In the first edition, in 2007 ([1], [2], [3]), the focus was mainly on the application to the Italian language of various parsing approaches, i.e. rule-based and statistical, and paradigms, i.e. constituency and dependency. The same development data, extracted from the Turin University Treebank (TUT[1]), have been therefore distributed both in dependency (TUT native) and constituency (TUT-Penn) format, and the Task was articulated in two parallel tracks, i.e. dependency and constituency. The results for dependency parsing have been evaluated as very close to the state of the art for English, while those for constituency showed a higher distance from it.

Notwithstanding the results of the Parsing Task in 2007 and the increasing popularity of the dependency parsing, the Evalita'09 Parsing Task includes again also a constituency parsing track, together with the track for dependency. While the development of new approaches for constituency parsing has allowed

[1] http://www.di.unito.it/~tutreeb

for the achievement of results of around F 92.1 [4] for English using the Penn Treebank, various experiences demonstrated that it is impossible to reproduce these results on other languages, e.g. Czech [5], German [6], Chinese [7], Italian [8], and on treebanks others than Penn [9]. Nevertheless, by proposing again the constituency track, we aim at contributing to the investigation on the causes of this irreproducibility with reference to Italian. Moreover, since the test set for the constituency and that for the dependency track share the same sentences, the contest makes available new materials for the development of cross-paradigm analyses about constituency parsing for Italian.

The focus of this paper is on the constituency track of the Evalita'09 Parsing Task only (see [10] for dependency track at the Evalita'09). The paper is organized as follows. In the following section, we describe the constituency track. In the other sections, we show the development data sets, the evaluation measures, the participation results and the discussion about them.

2 Definition of the constituency parsing task

The constituency parsing task is defined as the activity of assigning a syntactic structure to a given set of PoS tagged sentences (called *test set*), using a fully automatic parser and according to the constituency-based annotation scheme presented in a large set of sentences (called training or *development set*). The evaluation for this task is based on an annotated version of the test set (called *gold standard test set*).

3 Data sets

The data for the constituency track are from TUT, the treebank for Italian developed by the Natural Language Processing group of the Department of Computer Science of the University of Turin[2]. Even if smaller than other existing Italian resources, i.e. VIT [11] and ISST [12], TUT makes available more annotation formats [13], [14] that allowed for a larger variety of training and testing for parsing systems and for meaningfull comparisons with theoretical linguistic frameworks. In particular, since it is available both in dependency and constituency format, this treebank has been used in Evalita contests in 2007 and is used in 2009 as reference treebank. A recent project involving TUT concerns instead the development of the CCG-TUT, a treebank of Combinatory Categorial Grammar derivations for Italian [15].

The native annotation scheme of TUT is dependency-based, but in order to increase the comparability with other existing resources and in particular with the Penn Treebank, TUT has been converted in the TUT-Penn format, an adaptation of the Penn Treebank format for Italian. TUT-Penn has a morpho-syntactic tagging richer than the native Penn format used for English, but implements almost the same syntactic structure of Penn. In fact, since Italian is

[2] The free download of the resource is available at `http://www.di.unito.it/~tutreeb`.

inflectionally richer than English, TUT-Penn includes more PoS tags, i.e. 68 tags in TUT vs 36 in Penn Treebank [16]. Among the tags used in Penn, 22 are basic and 14 represent additional morphological features, while in TUT 19 are basic and 26 represent features. In both the tag sets, each feature can be composed with a limited number of basic tags, e.g. in TUT-Penn *DE* (for demonstrative) can be composed only with the basic tags *ADJ* (for adjective) and *PRO* (for pronoun). The higher number of tags in TUT-Penn is mainly motivated by the more fine-grained tagging for adjective, pronoun and especially verbs. For instance, in Penn a verb form can be annotated by using the basic tag *VB* (for basic form) or with a feature, e.g. *VBD* (for past tense) or *VBG* (for gerund or past participle), *VBN* (for past participle), *VBP* (non-third person singular present), *VBZ* (third person singular present), or can be annotated as *MD* (for modal). By contrary in TUT-Penn, the verb is annotated as *VMA* (for verb main), *VMO* (for modal) and *VAU* (for auxiliary) always associated with one of the 11 features that represent the conjugation, e.g. *VMO∼PA* for a modal verb in participle past or *VMA∼IM* for a main verb in imperfect.
With respect to the syntactic annotation of Penn Treebank, TUT-Penn is different only for some particular constructions and phenomena. For instance, TUT-Penn features a representation different from Penn for the subjects occurring after the verb. In Italian, where the word order is more free than in English, this is a quite common phenomenon, which is typically challenging for phrase structures (since the subject is considered as external argument of the VP). Therefore the TUT-Penn annotates a special functional tag, i.e. *EXTPSBJ* on the lexically realized subject (*PRO∼PE loro*) and a null element co-indexed with it, (*-NONE- *-233*) but positioned in the canonical position of the subject, as in the following example (sentence ALB-108 from the development set, subcorpus newspaper): 'Erano loro quelli che in città guadagnavano di più.' (word-by-word rough translation: *Were they that in the town gained more*).

```
( (S
   (NP-SBJ (-NONE- *-233))
   (VP (VMA~IM Erano)
       (NP-EXTPSBJ-233 (PRO~PE loro))
       (NP-PRD
           (NP (PRO~DE quelli))
           (SBAR
               (NP-4333 (PRO~RE che))
               (S
                   (NP-SBJ (-NONE- *-4333))
                   (PP-LOC (PREP in)
                       (NP (NOU~CA città)))
                    (VP (VMA~IM guadagnavano)
                        (ADVP (ADVB DI_PIÙ)))))))
         (. .)) )
```

Observe instead that TUT-Penn implements the same Penn annotation of null elements and coindexing, as you can see in the example where the relative

clause is represented as in the Penn format: the head of the structure is the relative pronoun (*PRO∼RE che*), and a null element coindexed with the pronoun (*-NONE- *-4333*) is the subject of the clause.
As well as TUT in native dependency format, the treebank in TUT-Penn format has been newly released in 2009 in an improved version where the annotation is more correct and consistent with respect to the previous version, that used in Evalita'07. Among the improvements of this last release there is the annotation of multi-word expressions, which are currently annotated by considering all the elements of an expression composed by more words as a single word or terminal node of the constituency tree (see e.g., in the previous example, the adverbial multi word expression 'di più' *more*).

The development set of the constituency track includes 2,200 sentences that correspond to 64,193 annotated tokens in TUT native format. The corpus is organized in two subcorpora, i.e. one from Italian newspaper (1,100 sentences and 33,522 tokens in TUT format) and one from the Italian Civil Law Code (1,100 sentences and 30,671 tokens in TUT format).

The test set includes 200 sentences and features the same balancement of the development set: 100 sentences from newspapers and 100 from the Civil Law Code. In order to make possible comparisons with the dependency track, the same sentences of the test set for the constituency track are included in the test set for the dependency track. More precisely, the dependency track includes two subtasks, i.e. the Main Subtask[3], whose test set includes all the 200 sentences of the test set for the constituency track, and the Pilot Subtask[4], whose test set includes the 100 sentences from newspapers of the test set for the constituency track.

4 Evaluation measures and participation results

We have used for the evaluation of constituency parsing results the standard metric EVALB ([5], `http://nlp.cs.nyu.edu/evalb/`): it is a bracket scoring program that reports labelled precision (**LP**), recall (**LR**), F-score (**LF**), non crossing and tagging accuracy for given data. Note that the official measure for the final result is the F-score. As usual we did not score the TOP label; moreover, in contrast with usual, in order to have a direct comparison with dependency subtask we do use punctuation in scoring.

We had two participants to the constituency track, i.e. FBKirst_Lavelli_CPAR and UniAmsterdam_Sangati_CPAR[5]. The parser from FBKirst adopts probabilistic context-free grammars model; the parser from University of Amsterdam adopts the DOP model.

[3] Where the development set includes, in TUT format, also the 2,200 sentences of the development set of the constituency track.
[4] Where the development set is extracted from ISST
[5] The name of each system that participated to the contest is composed according to the following pattern: institution_author_XPAR, where X is D for dependency and C for constituency.

The evaluation of the participation results for the constituency track is presented in table 1. Note that the difference between two results is taken to be significant if $p < 0.05$ (see `http://depparse.uvt.nl/depparse-wiki/AllScores` and `http://nextens.uvt.nl/~conll/software.html#eval`).
We can observe that the best results for this track have been achieved by the FBKirst_Lavelli_CPAR. Nevertheless, according to the p-value the difference between the first and second score cannot be considered as significant for recall.

Table 1. Constituency parsing: evaluation on all the test set (200 sentences).

LF	**LR**	**LP**	**p for LR**	**p for LP**	**Participant**
78.73	80.02	77.48	0.1592	0.0021	FBKirst_Lavelli_CPAR
75.79	78.53	73.24			UniAmsterdam_Sangati_CPAR

With respect to the subcorpora that represent the text genres of TUT within the test set, we can observe that the best results have been achieved on the set of sentences extracted from civil code section (see table 2).

Table 2. Constituency parsing: separate evaluation on the newspaper (100 sentences) and civil law (100 sentences) test set.

newspaper			**civillaw**			**Participant**
LF	**LR**	**LP**	**LF**	**LR**	**LP**	
76.21	76.08	76.34	80.66	83.15	78.33	FBKirst_Lavelli_DPAR
74.33	76.08	72.65	76.93	80.47	73.69	UniAmsterdam_Sangati_DPAR

5 Discussion

The results obtained in the constituency parsing track positively compare with the previous Evalita experience. The best results in 2007 and in the current contest have been achieved by the same team of research, i.e. FBKirst_Lavelli_DPAR, but with a different parser (Berkeley vs Bikel's [17]). In Evalita'07 the best scores are LF 67.97%, LP 65.36% and LR 70.81%, while in Evalita'09 the best scores are higher with an improvement of 10.76% for the LF, 12.12% for LP and 9.21% for LR.

Nevertheless, the distance from the results for English (F-score 92.1% [4]) remains high. Among the motivations for this distance it is important to take into account first of all the limited size of the treebank used for the training of the Italian parsers with respect to the Penn Treebank for English. Also the limited number of participants (two as in Evalita'07), which confirms the low popularity of the constituency parsing with respect to the dependency parsing for Italian (e.g. in Evalita'09 six participants for the main dependency subtask and five for the pilot), shows that we can expect from this area of research only a limited number of contributes and correlated evidences. Such a kind of evidences should come in the future also from comparisons among different constituency formats, as happened this year for dependency.

With respect to the text genres involved in the evaluation we can say that civil code is easier to parse with respect to newspaper texts in constituency parsing s well as in dependency [10].

Finally, with respect to dependency track, for constituency the results remain lower, but the distance is meaningfully lower than in 2007. It is well known in literature that is difficult to compare dependency and constituency evaluations, but the distance among the results can be clearly interpreted according to the hypothesis that dependency parsing is more adequate for Italian than the constituency parsing, at least with reference to the currently applied models and annotation schemes.

6 References

References

1. Bosco C., Mazzei A., Lombardo V.: Evalita Parsing Task: an analysis of the first parsing system contest for Italian. Intelligenza Artificiale, vol. 12, pp. 30–33 (2007)
2. Bosco, C., Mazzei, A., Lombardo, V., Attardi, G., Corazza, A., Lavelli, A., Lesmo, L., Satta, G., Simi, M.: Comparing Italian parsers on a common treebank: the Evalita experience. In: Proceedings of LREC'08, pp. 2066–2073 (2008)
3. Magnini, B., Cappelli, A., Tamburini, F., Bosco, C., Mazzei, A., Lombardo, V., Bertagna, F., Calzolari, N., Toral, A., Bartalesi Lenzi, V., Sprugnoli, R., Speranza, M.: Evaluation of Natural Language Tools for Italian: EVALITA 2007. In: Proceedings of LREC'08, pp. 2536-2543 (2008)
4. McClosky D., Charniak E., Johnson M.: When is self-training effettive for parsing? In: Proceedings of CoLing (2008)
5. Collins M., Hajic J., Ramshaw L., Tillmann C.: A Statistical Parser of Czech. In: Proceedings of ACL'99, pp. 505–512 (1999)
6. Dubey A., Keller F.: Probabilistic parsing for German using sister-head dependencies. In: Proceedings of ACL'03, pp. 96–103 (2003)
7. Levy R., Manning C.: Is it harder to parse Chinese, or the Chinese treebank? In: Proceedings of ACL'03, pp. 439–446 (2003)
8. Corazza A., Lavelli A., Satta G., Zanoli R.: Analyzing an Italian treebank with state-of-the-art statistical parser. In: Proceedings of TLT-2004, pp. 39–50 (2004)
9. Gildea D.: Corpus variation and parser performance. In: Proceedings of EMNLP'01, pp. 167–172 (2001)

10. Bosco C., Montemagni S., Mazzei A., Lombardo V., Dell'Orletta F., Lenci A.: Evalita'09 Parsing Task: comparing dependency parsers and treebanks. In: Proceedings of EVALITA 2009 (2009)
11. Delmonte R.: Strutture sintattiche dall'analisi computazionale di corpora di italiano. Franco Angeli, Milano (2008)
12. Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Lenci A., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M. T., Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R.: Building the Italian Syntactic-Semantic Treebank. In: Abeillé A. (ed.): Building and Using syntactically annotated corpora, pp. 189–210. Kluwer, Dordrecht (2003)
13. Bosco C., Lombardo V.: Comparing linguistic information in treebank annotations. In: Proceedings of LREC'06, pp. 1770–1775 (2006)
14. Bosco C.: Multiple-step treebank conversion: from dependency to Penn format. In: Proceedings of the Workshop on Linguistic Annotation at the ACL'07, pp. 164–167 (2007)
15. Bos, J., Bosco, C., Mazzei, A.: Converting a Dependency-Based Treebank to a Categorial Grammar Treebank for Italian. In: Proceedings of the 8th Workshop on Treebanks and Linguistic Theories, in press (2009)
16. Santorini B.: Part-of-Speech tagginf guidelines for the Penn Treebank project (3rd revision). Technical report no. MS-CIS-90-47 at University of Pennsylvania, http://repository.upenn.edu/cis_reports/570/
17. Corazza A., Lavelli A., Satta G.: Phrase Based Statistical Parsing. Intelligenza Artificiale, vol. 12, pp. 38–89 (2007)
18. Buchholz S., Marsi E.: CoNLL-X Shared Task on Multilingual Dependency Parsing. In: Proceedings of the CoNLL-X, pp. 149–164 (2006)
19. Nivre J., Hall J., Kübler S., McDonald R., Nilsson J., Riedel S., Yuret D.: The CoNLL 2007 Shared Task on Dependency Parsing. In: Proceedings of the EMNLP-CoNLL, pp. 915–932 (2007)