The Evalita 2011 Parsing Task: the constituency track

(Article begins on next page)

22 October 2019

The Evalita 2011 Parsing Task: the constituency track

Cristina Bosco and Alessandro Mazzei

Dipartimento di Informatica, Università di Torino
Corso Svizzera 185, 101049 Torino, Italy
{bosco,mazzei}@di.unito.it

Abstract. The aim of Evalita Parsing Task is at defining and extending
Italian state of the art parsing by encouraging the application of exist-
ing models and approaches. As in the Evalita'07 and '09, the Task is
organized around two tracks, i.e. Dependency Parsing and Constituency
Parsing. In this paper, we describe only the Constituency Parsing track
by presenting the data sets for development and testing, and reporting
the results, which positively compare with those obtained for this same
track held in the Evalita'07 and '09.

Keywords: Constituency Parsing, Evaluation, Italian

1 Motivation

The general aim of the Evalita Parsing evaluation campaign is at defining and
extending Italian state of the art parsing with reference to existing resources, by
encouraging the application of existing models to this language. As in previous
editions, in 2007 [7, 5] and 2009 [4, 3], the focus is mainly on the application to
Italian language of various parsing approaches, i.e. rule–based and statistical,
and paradigms, i.e. constituency and dependency–based. The aim is in fact at
contributing to the literature on parsing results giving information about the
behavior of parsing models on Italian, which is a morphologically rich language
currently less-resourced with respect e.g. to English or German.

In previous Evalita editions, the results for dependency parsing have been evalu-
ated as no far from the state of the art for English, while those for constituency
showed a higher distance from it. Instead, in the current edition the major im-
provement has to be referred to constituency parsing, where scores meaningfully
more proximate to the state of the art for English have been achieved. Never-
theless, these results confirm that the scores published for English (of around
F 92.1 [13]) using the Penn Treebank, remain currently irreproducible for Ital-
ian. By proposing again the constituency track, we aim at contributing to the
investigation on the causes of this irreproducibility and giving the same data
for development and testing (annotated in dependency and constituency for-
mat) in both tracks of the parsing task, we make available new materials for the
development of cross-paradigm analyses about Italian parsing.

In this paper we will analyze the constituency track of the competition, while the dependency track is described in [2]. The paper is organized as follows. In the following sections, first we describe the task, then, we show the development and test data sets and the measures applied in the evaluation procedure. We conclude with the presentation of participation results and a brief discussion.

2 Definition of the Task

As described in the CoNLL competitions [8, 14], the parsing task is defined as the activity of assigning a syntactic structure to a given set of PoS tagged sentences, called development set. A large set of syntactically fully annotated sentences is given to the participants in order to train and tune their parsers. The evaluation for this task is based on a manually syntactically annotated smaller set of sentences called gold standard test set.

Since the Evalita Parsing task is articulated in two tracks, for each track a test set and a development set is given by organizers. Moreover, in order to allow for a meaningful and direct comparison between the results achieved in the two tracks, all these datasets for dependency have been built by including exactly the same sentences as for constituency. Nevertheless, even if the organizers encouraged the participation to both tracks, only one participant submitted runs for both dependency and constituency.

3 Datasets

The data proposed for the training and development of parsing systems (i.e. the development set) 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[1]. TUT has been newly released for the last time in 2011, after automatic and manual revisions, in an improved version where both the consistency of the annotation and the size of the treebank are improved with respect to the previous releases. In particular, for what concerns size, TUT is currently similar to the other Italian resources, i.e. VIT and ISST–TANL. Moreover, TUT makes available more annotation formats [6] that allowed for a larger variety of training and testing for parsing systems and for meaningful comparisons with theoretical linguistic frameworks, e.g. the native TUT, the TUT-Penn, and the CCG-TUT which is an application to Italian of the Combinatory Categorial Grammar [1].

3.1 Development Set

For the constituency parsing track, the data format adopted is the TUT–Penn, as in previous Evalita contests, which is an application of the Penn Treebank format to the Italian language [3]. The kind and structure of the constituents

---

[1] For the free download of the resource, see `http://www.di.unito.it/∼tutreeb`.

are the same as in Penn Treebank for English, but the inventory of functional tags is enriched with some relations needed to represent e.g. the subject in post-verbal position. Moreover, in order to describe the rich inflectional system of Italian language, the TUT–Penn format adopts a different and richer set of Part of Speech tags with respect to the Penn Treebank.

The development set for both tracks includes the same sentences, namely 3,452 sentences belonging to the five different text genres which are currently represented in the subcorpora of the treeebank:

- NEWS and VEDCH, two collections of sentences from Italian newspaper (700 + 400 sentences)
- CODCIV, a collection of sentences from the Italian Civil Law Code (1,100 sentences)
- EUDIR, a collection of declarations of the European Community from the Italian section of the JRC-Acquis Multilingual Parallel Corpus[2] (201 sentences)
- Wikipedia, a collection of sentences from the Italian section of Wikipedia (459 sentences)
- COSTITA, the full collection of sentences of the *Costituzione Italiana* (682 sentences)

We can observe that in 2009 the development set for the constituency track was smaller and consisted in 2,200 sentences (i.e. 64,193 annotated tokens in TUT native format) representing only two text genres[3], which are now all included in the Evalita 2011 development set. Moreover, while in the past only a portion of the development set proposed for the training of dependency parsers was made available also for the constituency track in TUT–Penn format, this year, for the first time, the development set is exactly the same for both tracks.

3.2 Test Set

The test set for both the tracks is composed by the same 300 sentences which represent around the same balancement of the development set: 150 sentences from Civil Law Code, 75 sentences from newspapers and 75 sentences from Wikipedia. In Evalita 2009, the test set included only 200 sentences, namely 100 from newspapers and 100 from the Civil Law Code.

4 Evaluation Measures

As in previous editions of the contest, we exploited for the evaluation of constituency parsing results the standard metric EVALB (http://nlp.cs.nyu.edu/evalb/, [10]): it is a bracket scoring program that reports labelled precision (LP), recall

---

[2] http://langtech.jrc.it/JRC-Acquis.html

[3] The development set in 2009 was composed by the Italian newspapers (1,100 sentences) and the Civil Law Code (1,100 sentences) subcorpora.

(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 and the functional labels too; moreover, in contrast with usual, in order to have a direct comparison with dependency subtask we do use punctuation in scoring.

5   Participation Results

We had only one participant to the constituency track, i.e. FBKirst_Lavelli_CPAR[4], whose parser adopts probabilistic context-free grammars model, namely the Berkeley one. The author participated also to previous editions of the task, and to the dependency track of Evalita'11. Confirming the trend seen in previous editions of the task, the number of participants for constituency parsing has been smaller than for dependency, but in previous editions we had two participants also for the constituency track [3, 7].

The evaluation of the participation results for the constituency track is presented in table 1. With respect to the subcorpora that represent the text genres of TUT

Table 1. Constituency parsing evaluation on test set full (300 sentences).

Bracketing F	Bracketing Recall	Bracketing Precision	Participant
82.96	82.97	82.94	FBKirst_Lavelli_CPAR

within the test set, we can observe that the best results have been achieved on the set of sentences extracted from the Civil Law Code (see table 2).

Table 2. Constituency parsing evaluation on subcorpora.

Subcorpus	Size	Bracketing F	Bracketing Recall	Bracketing Precision
CODCIV	150	87.27	87.41	87.14
NEWS	75	77.46	78.22	76.72
WIKIPEDIA	75	78.38	77.49	79.30

---

[4] 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.

6 Discussion

The results obtained in the constituency parsing positively compare with previous experiences in this area, but still far from the state of the art for English constituency parsing.

Due to the fact that the task is based on the same (revised) treebank of Evalita'09 and Evalita'07, the more obvious comparison that we can develop is with this experience. With respect to the previous editions of the constituency parsing task in the Evalita'07 and '09, there is an impressive improvement of the results: the best F-score was in fact 72.15 in 2007, and 78.73 in 2009 (80.66 on the sentences from the Civill Law Code only) by Lavelli but by exploiting different parsers [9, 12].

As far as the text genre is concerned, we can observe that, in the dependency track of this year and in previous editions of the constituency parsing track, the best scores have been achieved on the Civil Law sentences. However, the differences between the performances of the participant parser on various genres are higher than in the case of dependency. Nevertheless, it is difficult to find the motivations of these differences because of the too limited participation to the track. We can only formulate two hypotheses. First, the parser adopted need for a larger amount of data to be trained successfully, and its performance on Civil Law is the best one because of the larger training set of data available. Second, the legal text genre is intrinsically less hard to parse with respect to newspapers or Wikipedia regardless of the amount of training data. Further experiments with other parsers and larger amount of data can help us in deciding if the first or second hypothesis applies.

References

1. 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, Milano (2009)
2. Bosco, C., Mazzei, A.: The Evalita 2011 Parsing Task: the dependency track. In Working Notes of EVALITA 2011, 24th-25th January 2012, Rome (2012)
3. Bosco, C., Mazzei, A., Lombardo, V.: Evalita Parsing Task 2009: constituency parsing and a Penn format for Italian. In: Proceedings of Evalita'09 at AI*IA, Reggio Emilia (2009)
4. 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'09 at AI*IA, Reggio Emilia (2009)
5. 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, Marrakesh (2008)
6. Bosco C.: Multiple-step treebank conversion: from dependency to Penn format. In: Proceedings of the Workshop on Linguistic Annotation at the ACL'07 (2007)
7. Bosco, C., Mazzei, A., Lombardo, V.: Evalita Parsing Task: an analysis of the first parsing system contest for Italian. Intelligenza Artificiale 12 (2007)

8. Buchholz S., Marsi E.: CoNLL-X Shared Task on Multilingual Dependency Parsing. In: Proceedings of the CoNLL-X (2006)
9. Corazza, A., Lavelli, A., Satta, G.: Phrase Based Statistical Parsing. Intelligenza Artificiale 12 (2007)
10. Collins M.: A New Statistical Parser Based on Bigram Lexical Dependencies. In: Proceedings of ACL96 (1996)
11. Kübler S., McDonald R., Nivre J.: Dependency parsing. Morgan and Claypool Publishers (2009)
12. Lavelli, A., Corazza, A.: The Berkeley Parser at the EVALITA 2009 Constituency Parsing Task. In: Proceedings of Evalita'09 at AI*IA, Reggio Emilia (2009)
13. McClosky D., Charniak E., Johnson M.: When is self-training effettive for parsing? In: Proceedings of CoLing (2008)
14. 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 (2007)