The Evalita 2011 Parsing Task: the Dependency 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 existing 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 Dependency Parsing track by presenting the data sets for development and testing, and reporting the test results.

Keywords: Dependency Parsing, Evaluation, Italian

1 Motivation

The Evalita Parsing evaluation campaign aims 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, which is morphologically rich and currently less-resourced. As in the editions held in 2007 [9, 7] and 2009 [3, 4], the focus is mainly on the application to Italian of various approaches, i.e. rule–based and statistical, and paradigms, i.e. constituency and dependency. Therefore, the task is articulated in two tracks, i.e. dependency and constituency, which share the same development and test data, distributed both in dependency and constituency format. In this paper we will analyze the dependency track of the competition, while the constituency track is described in [2].
The paper is organized as follows. First, we describe the task, then, we show the development and test data 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 [11, 17], the parsing task is the activity of assigning a syntactic structure to a given set of PoS tagged sentences. A large set of syntactically fully annotated sentences, i.e. the development set is given to the participants in order to train and tune their parsers. The evaluation is based on a manually syntactically annotated smaller set of sentences, called gold standard test set.

For each track in which the Evalita Parsing task is articulated, a test set and a development set is given by organizers. Moreover, in order to allow for a meaningful and direct comparison between 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 tracks.

3 Datasets

The data proposed for the training and development are from the Turin University Treebank (TUT), the treebank for Italian developed by the Natural Language Processing group of the Department of Computer Science of the University of Turin[1]. This resource, newly released in 2011, after automatic and manual revisions oriented to improving consistency and size of the treebank, is currently as large as other existing Italian resources, i.e. VIT and ISST–TANL. Moreover, in order to allow a variety of training and comparisons across various theoretical linguistic frameworks, TUT makes available several annotation formats [8], e.g. native TUT, TUT–Penn, and CCG–TUT, which is an application to Italian of the Combinatory Categorial Grammar [5].

The native scheme of TUT applies the major principles of dependency grammar [12] using a rich set of grammatical relations[2] and is featured by the distinctive inclusion of null elements to deal with non-projective structures, long distance dependencies, equi phenomena, pro drop and elliptical structures, which are quite common in a flexible word order language like Italian. On the one hand, this allows in the most of cases for the representation and the recovery of argument structures associated with verbs and nouns, and it permits the processing of long distance dependencies in a similar way to the Penn format. On the other hand, by using null elements crossing edges and non–projective dependency trees can be avoided.
Nevertheless, in order to make possible the application of standard evaluation measures within Evalita contests, the native format of TUT has been automatically converted in a format more proximate to the standard CoNLL. The resulting format differs from native TUT for the following features: it splits the annotation in the ten standard columns (filling eight of them) as in CoNLL, rather than organize them in round and square brackets; it exploits only part of the rich set of grammatical relations (72 in CoNLL versus 323 in TUT native); it does not include pointed indexes[3]. Since CoNLL does not allow null elements,

[1] For the free download of the resource, see `http://www.di.unito.it/~tutreeb`.
[2] See [10], [6].
[3] In TUT native format the representation of amalgamated words uses pointed indexes, e.g. a definite prepositions 'del' occurring as 33th word of a sentence is split in two lines, '33 del (PREP ....' and '33.1 del (ART ....' respectively representing the Preposition and the Article. In CoNLL format, where pointed indexes are not allowed, these two lines became '33 del (PREP ....' and '34 del (ART ....'.

they are deleted in this format, but the projectivity constraint is maintained at the cost of a loss of information with respect to native TUT in some cases.

3.1 Development Set

The development set includes 3,452 Italian sentences (i.e. 102,150 tokens in TUT native, and 93,987 in CoNLL[4]) and represents five different text genres organized in:

- NEWS and VEDCH, from newspapers (700 + 400 sentences, 18,044 tokens)
- CODCIV, from the Italian Civil Law Code (1,100 sentences, 28,048 tokens)
- EUDIR, from the JRC-Acquis Corpus[5] (201 sentences, 7,455 tokens)
- Wikipedia, from Wikipedia (459 sentences, 14,746 tokens)
- COSTITA, the full text of the *Costituzione Italiana* (682 sentences, 13,178 tokens)

In 2009 the development set was smaller, and consisted in 2,400 sentences (i.e. 72,149 annotated tokens in TUT native format, and 66,055 in CoNLL format), organized in only three subcorpora, i.e. Italian newspapers, Civil Law Code and JRC-Acquis. All these sentences are included also in the Evalita 2011 development set.

3.2 Test Set

The test set is composed by 300 sentences (i.e. 7,836 tokens) around balanced as in the development set: 150 sentences from Civil Law Code (3,874 tokens), 75 sentences from newspapers (2,035 tokens) and 75 sentences from Wikipedia (1,927 tokens). In Evalita 2009 the test set included 240 sentences (5,287 tokens).

4 Evaluation Measures

The standard methodology for the evaluation of dependency parsers is to apply them to a test set and compare their output to the gold standard test set, i.e. the test set annotated according to the treebank used for the development of the parsers.
Among the most widely used evaluation metrics, we have selected for the evaluation of the official results those used in the CoNLL parsing shared task, i.e. LAS (Labeled Attachment Score) that is the percentage of tokens with correct head and dependency type. Moreover, in accord with literature, we report the UAS (Unlabeled Attachment Score) measure too, i.e. the percentage of tokens with correct head [11, 17]. Note that the use of a single accuracy metric is possible in dependency parsing thanks to the single-head property of dependency trees, which implies that the amount n of nodes/words always corresponds to $n - 1$ dependency relations. This property allows the unification of measures of precision and recall and makes parsing resemble a tagging task, where every word is to be tagged with its correct head and dependency type [13].

[4] In the following we will refer only to number of tokens in CoNLL format.
[5] `http://langtech.jrc.it/JRC-Acquis.html`

5 Participation Results

The participants[6] to the dependency parsing track were four. Among them only one did not participate at the previous editions of the contest, where the participants were six.

Two participant systems, i.e. UniTo_Lesmo_DPAR and Parsit_Grella_DPAR, do not follow the statical approach. UniTo_Lesmo_DPAR system is a rule–based wide coverage parser developed in parallel with TUT, which has been applied to various domains. The Parsit_Grella_DPAR uses instead a hybrid approach that mixes rules and constraints. The other two participating systems belong instead to the class of statistical parsers: FBKirst_Lavelli_DPAR is an application to Italian of different parsing algorithms implemented in MaltParser [16] and of an ensemble model made available by Mihai Surdeanu; UniPi_Attardi_DPAR is instead DeSR, a Shift/Reduce deterministic transition–based parser [1] which participated also to CoNLL contests.

According to the main evaluation measure, i.e. LAS, the best results have been achieved by Parsit_Grella_DPAR followed by UniPi_Attardi_DPAR (see table 1) with a difference statistically significant according to the p-value[7]. The average scores of the participants are 88.76 for LAS and 93.55 for UAS.

Table 1. Dependency parsing: evaluation on the test set (300 sentences).

LAS	**UAS**	**Participant**
91.23	96.16	Parsit_Grella_DPAR
89.88	93.73	UniPi_Attardi_DPAR
88.62	92.85	FBKirst_Lavelli_DPAR
85.34	91.47	UniTo_Lesmo_DPAR

In table 2, we see instead how the performance varies according to text genres. If evaluated on the civil law texts the difference among the three best scored systems is not statistically significative, while it is significative on Wikipedia and more valuable on newspaper. In the latter text genre, all the scores achieved by Parsit_Grella_DPAR are significantly higher than those of the others, and this motivates the success of this parser in the contest.

[6] 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.

[7] 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`).

Table 2. Dependency parsing, evaluation on subcorpora: civil law (150 sentences), newspaper (75 sentences), wikipedia (75 sentences).

civillaw		**newspaper**		**wikipedia**		**Participant**
LAS	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**	
92.85	96.18	86.34	91.19	86.91	90.88	UniPi_Attardi_DPAR
92.21	97.01	90.75	95.54	89.51	94.51	Parsit_Grella_DPAR
91.56	95.12	83.84	89.72	87.09	91.05	FBKirst_Lavelli_DPAR
89.06	94.43	80.69	87.70	81.87	88.80	UniTo_Lesmo_DPAR

6 Discussion

The results positively compare with other experiences, both for Italian, i.e. Evalita'07 and '09, which were based on the same (revised) treebank, and for other languages, e.g. English (LAS $89,61\%$) and Japanese (LAS $91,65\%$) [17]. The best scores passed from 86.94 for LAS and 90.90 for UAS in 2007 (by UniTo_Lesmo_DPAR), to 88.73–88.69 for LAS(by UniTo_Lesmo_DPAR and UniPi_Attardi_DPAR) and 92.72 for UAS (by UniPi_Attardi_DPAR) in 2009, to 91.23 for LAS and 96.16 for UAS by Parsit_Grella_DPAR in 2011. The average LAS is passed from 72.48 in 2007, to 82.88 in 2009), to 88.76 in 2011, while the average UAS from 83.09, to 87.96, to 93.55. As in previous editions, the best performances are referred to the Civil Law texts.
With respect to the approaches, the observed improvement of results for the two statistical systems can be probably motivated by the availability of larger sets of data for training from 2007 to 2011. For instance, by contrast with Evalita'09 results, the top rule–based parser in Evalita'09 and '07 (UniTo_Lesmo_DPAR) scores significantly worst than the two stochastic parsers (UniPi_Attardi_DPAR) and FBKirst_Lavelli_DPAR. But the best performing system is again a non statistical system (i.e. Parsit_Grella_DPAR). Nevertheless, also the results of this edition confirm that non–statistical systems can achieve good scores only if developed in parallel with the reference resource, like UniTo_Lesmo_DPAR in the past contests and Parsit_Grella_DPAR; while rule–based approaches not enough tuned on the resource obtained negative results, see e.g. [19] or [18].
We conclude by observing that in a wider perspective of evaluation of the contribution of parsing to the overall quality of applicative NLP systems, other kinds of information should be taken into account, in particular those coming from null elements and semantic features currently annotated only in a few resources.

References

1. Attardi, G., DellOrletta, F., Simi, M., Turian J.: Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In: Proceedings of Evalita'09 at AI*IA,

Reggio Emilia (2009)
2. Bosco, C., Mazzei, A.: The Evalita 2011 Parsing Task: the constituency track. In Working Notes of EVALITA 2011, 24th-25th January 2012, Rome (2012)
3. 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)
4. Bosco, C., Mazzei, A., Lombardo V.: Evalita'09 Parsing Task: constituency parsers and the Penn format for Italian. In: Proceedings of Evalita'09 at AI*IA, Reggio Emilia (2009)
5. 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)
6. Bosco C.: A grammatical relation system for treebank annotation. Unpublished PhD thesis discussed at the University of Turin (2004)
7. 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)
8. Bosco C.: Multiple-step treebank conversion: from dependency to Penn format. In: Proceedings of the Workshop on Linguistic Annotation at the ACL'07, Prague (2007)
9. Bosco, C., Mazzei, A., Lombardo, V.: Evalita Parsing Task: an analysis of the first parsing system contest for Italian. Intelligenza Artificiale 12 (2007)
10. Bosco C., Lombardo V., Vassallo D., Lesmo L.: Building a Treebank for Italian: a Data-driven Annotation Schema. In: Proceedings of LREC'00, Athens, Greece (2000)
11. Buchholz S., Marsi E.: CoNLL-X Shared Task on Multilingual Dependency Parsing. In: Proceedings of the CoNLL-X (2006)
12. Hudson R.: Word grammar. Basil Blackwel, Oxford and New York (1984)
13. Kübler S., McDonald R., Nivre J.: Dependency parsing. Morgan and Claypool Publishers (2009)
14. Levy R., Manning C.: Is it harder to parse Chinese, or the Chinese treebank? In: Proceedings of ACL'03 (2003)
15. 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, Marrakesh (2008)
16. Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S., Marsi E.: MaltParser: a language-independent system for data-driven dependency parsing. In: Natural Language Engineering 13(2) (2007)
17. 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)
18. Testa, M., Bolioli, A., Dini, L., Mazzini, G.: Evaluation of a Semantically Oriented Dependency Grammar for Italian at EVALITA 2009. In: Proceedings of Evalita'09 at AI*IA, Reggio Emilia (2009)
19. Zanzotto, F.M.: Lost in Grammar Translation. Intelligenza Artificiale 12 (2007)