EVALITA 2011

Italian Parsing Evaluation

Task Guidelines

Cristina Bosco, Alessandro Mazzei
Dipartimento di Informatica, Università di Torino, Italy
{bosco, mazzei}@di.unito.it

1. Introduction

The following are the guidelines for the parsing task of the EVALITA 2011 evaluation campaign.

The participation is open to parsing systems pursuing different approaches, e.g. rule-based versus statistical.

The task includes two tracks, i.e. Dependency Parsing (see 1.1) and Constituency Parsing (see 1.2). All participants can participate to a single track, but are strongly encouraged to perform both the tracks.

The data for the two tracks (see 2.) are made available in both a dependency-based (see 3.1) and a constituency-based format (see 3.2). The evaluation will be performed according to the standard metrics known in literature (see 4.).

Participants are required to provide a brief description of their system, an illustration of their experiments, in particular techniques and resources used, and an analysis of their results for the publication in the Proceedings of contest (see `http://www.evalita.it/2011/proceedings` for updated information about the publication of proceedings).

1.1 Dependency Parsing

The Dependency Parsing track (**DepPar**) is based on the Turin University Treebank (TUT), developed by the University of Torino by semi-automatic tools (`http://www.di.unito.it/~tutreeb`) [1, 2, 3], and also used as the reference treebank for dependency parsing in the previous edition of Evalita in 2007 [4, 5, 6] and 2009 [7, 8].

The evaluation will exploit two data sets provided by the organizers: the first, referred to as **DevSet-DepPar**, contains data annotated using the TUT format (see 3.1) and must be used for the development and training of the systems participating to the dependency track; the second, referred to as **TestSet-DepPar**, contains blind test data PoS-tagged according to the TUT PoS tag set.

1.2 Constituency Parsing

The Constituency Parsing track (**CosPar**) is based on the TUT-Penn treebank. This treebank is the result of the application to the TUT of a fully automatic conversion implemented at the University of Torino. As well as TUT for dependency parsing, TUT-Penn has been the reference treebank for constituency parsing in the previous Evalita editions in 2007 [4, 5, 6] and 2009 [9].

For the CosPar, the evaluation will exploit two data sets provided by the organizers: the first, referred to as **DevSet-CosPar**, contains data annotated using the TUT-Penn format [10, 11, 12] (see 3.2) and must be used for the development and training of the constituency track participating systems; the second, referred to as **TestSet-CosPar**, contains blind test data PoS-tagged according to the TUT-Penn tag set[1].

2. Corpora Description

As said above, for each track, a development set will be provided by the organizers, i.e. the DevSet-DepPar and the DevSet-CosPar, respectively for the Dependency and the Constituency track. The data are the same in both sets, but annotated according to the dependency format of the TUT (see 3.1) in the DevSet-DepPar, and the TUT-Penn format (see 3.2) in the DevSet-CosPar. As in the previous Evalita editions, this is done in order to guarantee the comparison of the DepPar track with the CosPar track.

All TUT and TUT-Penn materials are covered by a Creative Commons license for free software and are available for download from the treebank web site.

With respect to the previous edition of the contest, the DevSets has been quantitatively and qualitatively improved: on the one hand, they have been extended by including new corpora developed in the last years within the TUT project, i.e. WIKI and COSTITA; on the other, all data have been manually and automatically checked and newly released for the Evalita 2011 parsing task.

The data of the DevSets are organized in the following corpora:

- NEWS, which is composed of sentences coming from various newspapers, the most concerning the situation in Albania in the nineties;

- VEDCH, which contains sentences that include the Italian verb "vedere" (to see; 350 sentences) and "chiamare" (to call; 50 sentences);

- CODICECIVILE, which includes sentences extracted from the Italian Civil Code of law;

- EUDIR, which contains sentences from the Italian section of the JRC-Aquis multilingual parallel corpus[2];

- WIKI, which includes sentences randomly chosen from the Italian version of Wikipedia;

- COSTITA, which includes all the sentences of the Costituzione Italiana.

---

[1] In order to guarantee the comparison of results of the Dependency and Constituency Parsing, the unannotated data of the TestSet-CosPar are the same as TestSet-DepPar, but their morpho-syntactic annotation varies since TUT and TUT-Penn exploits different tag sets.

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

The table below shows the distribution of data in corpora and the size of each corpus in terms of tokens annotated in CoNLL format (see 3.1).

corpus	sentences	tokens (in CoNLL)
NEWS	700	18,044
VEDCH	400	12,508
CODICECIVILE	1,100	28,048
EUDIR	201	7,426
WIKI	459	14,746
COSTITA	682	13,860
all	3,542	94,632

3. Data Formats

The DevSets and the TestSets will be provided as Unix UTF-8 encoded files, for Dependency Parsing DevSet-DepPar and TestSet-DepPar, and for Constituency Parsing DevSet-CosPar and TestSet-CosPar.

3.1. Dependency Parsing Data Formats

In the DevSet-DepPar and in the TestSet-DepPar there is one token per line followed by its tags, separated by a TAB and organized according to the standard below described CoNLL 10-colums format [13]:

Field number:	Field name:	Description:
1	ID	Token counter, starting at 1 for each new sentence.
2	FORM	Word form or punctuation symbol.
3	LEMMA	Lemma, or _ if not available.
4	CPOSTAG	Coarse-grained part-of-speech tag.
5	POSTAG	Fine-grained part-of-speech tag, for TUT identical to the coarse-grained part-of-speech tag.
6	FEATS	Set of syntactic and/or morphological features, separated by a |, or _ if not available.
7	HEAD	Head of the current token, which is either a value of ID or zero ('0'). There cannot be multiple tokens with an ID of zero.
8	DEPREL	Dependency relation to the HEAD. The set of dependency relations is that of TUT/TANL.
9	PHEAD	Projective head, or _ if not available.
10	PDEPREL	Projective dependency relation, or _ if not available.

An empty line terminates each sentence.

The annotation scheme applied in the data is that of TUT, which is shown in the following example as represented according to the CoNLL format, one token per line[3]. For a more detailed description of the adopted annotation scheme, see TUT web site and/or the referred literature, where all the data of the DevSet-DepPar are also available in TUT native format.

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL	PHEAD	PDEPREL
1	Il	IL	ART	ART	DEF|M|SING	7	RMOD	_	_
2	19	>19>	NUM	NUM	_	1	ARG	_	_
3	novembre	NOVEMBRE	NOUN	NOUN	COMMON|M|ALLVAL	2	RMOD	_	_
4	i	IL	ART	ART	DEF|M|PL	7	OBJ/SUBJ	_	_
5	berlinesi	BERLINESE	NOUN	NOUN	COMMON|ALLVAL|PL	4	ARG	_	_
6	saranno	ESSERE	VERB	VERB	AUX|IND|FUT|INTRANS|3|PL	7	AUX+PASSIVE	_	_
7	chiamati	CHIAMARE	VERB	VERB	MAIN|PARTICIPLE|PAST|TRANS|PL|M	0	TOP	_	_
8	a	A	PREP	PREP	MONO	7	INDCOMPL	_	_
9	manifestare	MANIFESTARE	VERB	VERB	MAIN|INFINITE|PRES|TRANS	8	ARG	_	_
10	per	PER	PREP	PREP	MONO	9	RMOD	_	_
11	la	IL	ART	ART	DEF|F|SING	10	ARG	_	_
12	libertà	LIBERTÀ	NOUN	NOUN	COMMON|F|ALLVAL	11	ARG	_	_
13	di	DI	PREP	PREP	MONO	12	RMOD	_	_
14	stampa	STAMPA	NOUN	NOUN	COMMON|F|SING	13	ARG	_	_
15	.	#\.	PUNCT	PUNCT	_	7	END	_	_

The version of the data of the DevSet-DepPar includes for each token a single line annotated by using the morpho-syntactic information and the syntactic relations of the TUT format. Instead, the blind version of the data for the test set, i.e. TestSet-DepPar, will only contain morpho-syntactically annotated tokens, one per line. Therefore, to be more concrete, the development data contain the first eight columns annotated, while the test data will contain only the first six columns.

The format of the run files submitted by participants must be the same as in the DevSet-DepPar, i.e. containing one token per line including all original columns of

---

[3] The TUT native format is richer than that described in these guidelines, both for the larger set of grammatical relations and the use of null elements. While the set of grammatical relations has been reduced for limiting the sparseness of data, null elements have been avoided since they are not allowed in CoNLL standard format.

the test data of the TesSet-DepPar plus the two columns for the syntactic annotation, i.e. the HEAD and DEPREL columns.

3.2.Constituency Parsing Data Format

The development corpus for Constituency Parsing, DevSet-CosPar, is one sentence for each line, as required by the EVALB evaluation metrics [14]. The words are organized as in Penn Treebank format for what concerns the phrase structure of the sentence, like in the following example.


```
( (S
    (NP-SBJ (-NONE- *-233))
    (VP (VMA~RE Piovono)
            (NP-EXTPSBJ-233
                (NP (NOU~CP pietre))
                (CONJ e)
                (NP (NOU~CP insulti)))
        (, ,)
        (PP (ADVB anche)
            (PREP contro)
            (NP
                (NP (ART~DE gli) (NOU~CP stranieri))
                (CONJ e)
                (NP (ART~DE gli) (NOU~CP italiani)))))
    (. .)) )
```


It can be observed that, in order to better describe the rich inflection of Italian, the PoS tagging of TUT-Penn is different from that in Penn Treebank (see the TUT web site for a more detailed description of the PoS and functional tags adopted for TUT-Penn). Instead, the structure and the functional relations, which describe the constituency trees, are around the same as in Penn Treebank.

The blind version of the data for the test set, i.e. TestSet-CosPar, will contain just non syntactically annotated tokens, one word for each line, as in the following example:

1 Piovono (VMA~RE)

2 pietre (NOU~CP)

3 e (CONJ)

4 insulti (NOU~CP)

5 , (PUNCT)

6 anche (ADVB)

7 contro (PREP)

8 gli (ART~DE)

9 stranieri (NOU~CP)

10 e (CONJ)

11 gli (ART~DE)

12 italiani (NOU~CP)

13 . (PUNCT)

The format of the submitted run files must be the same as DevSet-CosPar, with the corresponding TUT-Penn functional tags and phrase structure.

4. **Evaluation Metrics**

For the Dependency Parsing track, the evaluation metrics are labelled and unlabeled attachment score, like in CoNLL shared tasks: LAS is the proportion of "scoring" tokens that are assigned both the correct head and the correct dependency relation label, whilst UAS is the proportion of "scoring" tokens that are assigned the correct head (regardless of the dependency relation label) [13].

For the Constituency Parsing track, the evaluation metric is instead tree precision and recall no-crossing bracket metric calculated by using the EVALB program [14].

5. **Evaluation Details**

For all the evaluation details and important dates, the participants should refer to the Evalita web page (`http://www.evalita.it/2011`) where the information will be updated all during the Evalita campaign.

Participants should submit the results of their runs sending, to the organizers email address (bosco@di.unito.it), one file for each track they want to participate to, in the same format as the relative development corpus, as said above. Each file has to be named as:

EVALITA11_PAR_TRACK_Org-Participantname

where TRACK has to be substituted by the name of the track (COS for Constituency Parsing and DEP for Dependency Parsing), and Org-Participantname has to be substituted by the name of the organization and the surname of the participant, e.g. "EVALITA11_PAR_DEP_UniTo-Bosco" should be the name for the file of Bosco of the University of Torino for the dependency track.
Observe that only results in the right format and one result for each track will be accepted for each participant.
After the submission deadline, the organizers will evaluate the submitted runs and will send each participant the score of his submissions as well as the gold-standard version of the test sets.

6.References

[1] C. Bosco, V. Lombardo, D. Vassallo, L. Lesmo (2000). Building a treebank for Italian: a data-driven annotation schema. In Proceedings of LREC 2000, Athens.

[2] C. Bosco, V. Lombardo (2003). A relation-schema for treebank annotation. In Proceedings of AI*IA 2003, Pisa.

[3] L. Lesmo, V. Lombardo, C. Bosco (2002). Treebank Development: the TUT Approach. In Proceedings of ICON 2002, Mumbay.

[4] C. Bosco, A. Mazzei, V. Lombardo (2007). Evalita Parsing Task: an analysis of the first parsing system contest for Italian. Intelligenza artificiale, anno IV, num 2, June 2007.

[5] C. Bosco, A. Mazzei, V. Lombardo, G. Attardi, A. Corazza, A. Lavelli, L. Lesmo, G. Satta, M. Simi (2008). Comparing Italian parsers on a common treebank: the Evalita experience. In Proceedings of LREC'08, Marrakesh.

[6] B. Magnini, A. Cappelli, F. Tamburini, C. Bosco, A. Mazzei, V. Lombardo, F. Bertagna, N. Calzolari, A. Toral, V. Bartalesi Lenzi, R. Sprugnoli, M. Speranza (2008). Evaluation of Natural Language Tools for Italian: EVALITA 2007. In Proceedings of LREC'08, Marrakesh.

[7] C. Bosco, S. Montemagni, A. Mazzei, V. Lombardo, F. Dell'Orletta, A. Lenci (2009). Evalita'09 parsing task: comparing dependency parsers and treebanks. In Proceedings of Evalita'09, Reggio Emilia.

[8] C. Bosco, S. Montemagni, A. Mazzei, V. Lombardo, F. Dell'Orletta, A. Lenci, L. Lesmo, G. Attardi, M. Simi, A. Lavelli, J. Hall, J. Nilsson, J. Nivre (2010). Comparing the influence of different treebank annotations on dependency parsing. In Proceedings of LREC'10, Malta.

[9] C. Bosco, A. Mazzei, V. Lombardo (2009). Evalita'09 parsing task: constituency parsers and the Penn format for italian. In Proceedings of Evalita'09, Reggio Emilia.

[10] C. Bosco (2007). Multiple-step treebank conversion: from dependency to Penn format. In Proceedings of Linguistic Annotation Workshop (LAW) at ACL'07, Prague.

[11] C. Bosco (2006). Linguistic knowledge extraction from corpus parallel annotations. In Proceedings of XL Congresso della Società di Linguistica Italiana, Vercelli.

[12] C. Bosco, V. Lombardo (2006). Comparing linguistic information in treebank annotations. In Proceedings of LREC'06, Genova.

[13] J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, D. Yuret (2007). The CoNLL 2007 Shared Task on dependency parsing. In Proceedings of EMNLP-CoNLL'07, Prague.

[14] M. Collins (1996). A New Statistical Parser Based on Bigram Lexical Dependencies. In Proceedings of ACL'96, San Francisco.