EVALITA 2009

Italian Parsing Evaluation

Task Guidelines

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

Felice dell'Orletta[1], Alessandro Lenci[2], Simonetta Montemagni[3]
[1]Università di Pisa, Dipartimento di Informatica
[2]Università di Pisa, Dipartimento di Linguistica
[3]Istituto di Linguistica Computazionale (ILC) - CNR
{felice.dellorletta, alessandro.lenci, simonetta.montemagni}@ilc.cnr.it

1. Introduction

The following are the guidelines for the parsing task of the EVALITA 2009 evaluation campaign.

The task includes two tracks, i.e. Dependency Parsing and Constituency Parsing. The participation is open to parsing systems pursuing different approaches, e.g. rule-based vs statistical.

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.

1.1 Dependency Parsing

The Dependency Parsing track will be articulated into two subtasks:

- The **main dependency subtask** (**mainDepPar**) uses as the development set the corpora of 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].

  For the mainDepPar the evaluation will be based on two data sets provided by the organizers: the first, referred to as **DevSet-mainDepPar**, contains data annotated using the *TUT format* and must be used for the development and training of all dependency track participating systems; the second, referred to as **TestSet-mainDepPar**, contains blind test data PoS-tagged according to the TUT PoS tag set.

- The **pilot dependency subtask** (**pilotDepPar**) uses as the development set the TANL dependency annotated corpus jointly developed by the Istituto di Linguistica Computazionale (ILC-CNR) and the University of Pisa in the framework of the project “Analisi di Testi per il Semantic Web e il Question Answering” [7]. The TANL dependency annotated corpus originates as a revision of the ISST-CoNLL corpus [8], built in its turn starting from the Italian Syntactic-Semantic Treebank [9]; in particular, the ISST-CoNLL corpus was built on top of the ISST morpho-syntactic and syntactic dependency annotation levels through a semi-automatic conversion process in charge of

a) combining information coming from two different annotation levels, and b) converting the ISST annotation scheme for dependency annotation into the CoNLL-2007 tabular format.

For the pilotDepPar, the evaluation will be based on two data sets provided by the organizers: the first, referred to as **DevSet-pilotDepPar**, contains data annotated using the *TANL format* and must be used for the development and training of the pilot subtask participating systems; the second, referred to as **TestSet-pilotDepPar**, contains blind test data PoS-tagged according to the TANL PoS tag set[1].

The mainDepPar is the obligatory subtask for all the participants to the Dependency Parsing track, while the pilotDepPar is an optional subtask. Nevertheless, all participants are strongly encouraged to perform both mainDepPar and pilotDepPar tasks.

1.2 Constituency Parsing

The Constituency Parsing track (**CosPar**) will consist in a single track, which uses as the development set 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 edition [4, 5, 6].

For the CosPar, the evaluation will be based on 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] and must be used for the development and training of all Constituency Parsing track participating systems; the second, referred to as **TestSetCP**, contains blind test data PoS-tagged according to the TUT-Penn tag set[2].

2. Corpora Description

Three data sets will be provided by the organizers:

- The development corpus for the **mainDepPar** is **DevSet-mainDepPar**, which includes two corpora and, in particular:
    - the <u>TUT original corpus</u> that currently consists in 2,200 sentences (64,215 tokens in dependency annotation), 1,100 (33,534 tokens) extracted from Italian newspapers and 1,100 (30,681 tokens) from Italian Civil Law Code; starting from the Evalita 2007 Parsing Task development and test sets, this corpus has been newly released in March 2009 in an extended and improved version where the annotation has been automatically and manually revised in order to increase correctness and consistency of the data.

[1] Note that in order to guarantee the comparison of results achieved on the basis of different developments sets (differing in size, composition, granularity and annotation schemes) there will be an intersection between the unannotated data of **TestSet-mainDeP** and **TestSet-pilotDeP**.

[2] 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-mainDeP**, but their morpho-syntactic annotation varies since TUT and TUT-Penn exploits different tag sets.

- the JRC-Passage-Evalita corpus that consists in 200 sentences (around 2,800 tokens in dependency annotation) extracted from the parallel JRC-Acquis Multilingual Parallel Corpus (http://langtech.jrc.it/JRC-Acquis.html). The Italian version of this corpus has been annotated, according to the TUT format, for Evalita 2009, and the French version has been annotated, according to the Easy format, for Passage 2009 (http://atoll.inria.fr/passage/index.en.html), which is an evaluation campaign on parsing for French language.

On the one hand, the TUT corpus guarantees the comparison of the Dependency Parsing track with the Constituency Parsing track; in fact, the corpus DevSet-CosPar for the CosPar, is this same TUT corpus, but annotated according to the TUT-Penn format.
On the other hand, the JRC-Passage-Evalita corpus permits preliminary comparisons with Easy and French language, since this small corpus will be included in both the development set of Passage and Evalita.
All TUT materials are covered by a license for free software and are available for download from the treebank web site.

- For the **pilotDepPar**, the **DevSet-pilotDepPar** corpus has been randomly extracted from the TANL dependency annotated corpus for a total of 3,109 sentences and 71,273 tokens. This corpus is a subset of the balanced partition of the ISST corpus exemplifying general language usage and consisting of a selection of articles from newspapers and periodicals, all selected to cover a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.). The TANL dependency annotated corpus is a revised version of the ISST-CoNLL corpus, where revisions – all performed manually - were mainly concerned with the adopted dependency Tag Set and the annotation criteria. The corpus is copyrighted material which can be used for research purposes only and which cannot be distributed in any original or modified form (see the licence agreement form).

  Note that the pilot subtask, based on a revised version of the ISST-CoNLL corpus already used for Italian in the multi-lingual track of CoNLL 2007 Shared Task on Dependency Parsing, creates the prerequisites for comparing EVALITA-2009 results with state-of-the-art dependency parsing at CoNLL-2007.

- In order to guarantee the comparison of the Dependency and Constituency Parsing tracks, as said before, for the **CosPar** the development set **DevSet-CosPar** is the TUT original corpus, but annotated according to the TUT-Penn format. This corpus currently consists in 2,200 sentences (64,215 tokens in dependency annotation), 1,100 (33,534 tokens) extracted from Italian newspapers and 1,100 (30,681 tokens) from Italian Civil Law Code. Also TUT-Penn has been newly released in March 2009 in a version that extends and improves the development and test sets of the Evalita 2007 Parsing Task, where the annotation has been automatically and manually revised in order to increase correctness and consistency of the data. All TUT-Penn materials are covered by a license for free software and are available for download from the treebank web site.

It is worth emphasising here that comparison of results achieved on different corpora will result in interesting insights into whether and how the annotation features of a treebank can influence the parser performance. Starting from this comparison we are planning to create a larger unified resource for Italian in which individual dependency annotated corpora will be combined together and which will hopefully be used for the next EVALITA editions.

3. Data Formats

The development corpora will be provided as three Unix file, two for Dependency Parsing, i.e. **DevSet-mainDepPar** and **DevSet-pilotDepPar**, and one for Constituency Parsing, i.e. **DevSet-CosPar**.

3.1.Dependency Parsing Data Formats

Both the development corpora for Dependency Parsing are UTF-8 encoded, one token per line followed by its tags, separated by a TAB and organized according to the CoNLL 10-colums format:

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.

In spite of the fact that both corpora adhere to this same format, they differ in size, composition and annotation schemes. Size and composition of the two corpora is described in Section 2 above. Annotation scheme differences are concerned with both Tag Sets and annotation criteria, as it can be noticed by comparing the sentences below annotated respectively according to the TUT and the TANL formats:

- Example sentence from TUT in CoNLL 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|	7	AUX+PASSIVE	_	_

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL	PHEAD	PDEPREL
					INTRANS|3|PL				
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	_	_

- Example sentence from the TANL dependency annotated corpus:

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL	PHEAD	PDEPREL
1	Per	Per	E	E	_	8	comp_temp	_	_
2	tutta	Tutto	T	T	num=s|gen=f	4	mod	_	_
3	la	Lo	R	RD	num=s|gen=f	4	det	_	_
4	giornata	giornata	S	S	num=s|gen=f	1	prep	_	_
5	i	Il	R	RD	num=p|gen=m	6	det	_	_
6	carabinieri	carabiniere	S	S	num=p|gen=m	8	subj	_	_
7	hanno	Avere	V	VA	num=p|per=3|mod=i|ten=p	8	aux	_	_
8	controllato	controllare	V	V	num=s|mod=p|gen=m	0	ROOT	_	_
9	decine	decina	S	S	num=p|gen=f	8	obj	_	_
10	di	Di	E	E	_	9	comp	_	_
11	persone	persona	S	S	num=p|gen=f	10	prep	_	_
12	,	,	F	FF	_	17	punc	_	_
13	tra	Tra	E	E	_	17	comp	_	_
14	cui	Cui	P	PR	num=n|gen=n	13	prep	_	_
15	i	Il	R	RD	num=p|gen=m	17	det	_	_
16	cinque	cinque	N	N	_	17	mod	_	_
17	utilizzatori	utilizzatore	S	S	num=p|gen=m	11	mod_rel	_	_
18	del	Di	E	EA	num=s|gen=m	17	comp	_	_
19	box	Box	S	S	num=n|gen=m	18	prep	_	_
20	dove	Dove	B	B	_	22	mod_loc	_	_
21	sarebbe	essere	V	VA	num=s|per=3|mod=d|ten=p	22	aux	_	_
22	avvenuta	avvenire	V	V	num=s|mod=p|gen=f	19	mod_rel	_	_
23	la	Lo	R	RD	num=s|gen=f	24	det	_	_
24	violenza	violenza	S	S	num=s|gen=f	22	subj	_	_
25	:	:	F	FC	_	8	punc	_	_

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL	PHEAD	PDEPREL
26	box	Box	S	S	num=n|gen=m	19	mod	_	_
27	che	Che	P	PR	num=n|gen=n	30	subj_pass	_	_
28	è	essere	V	VA	num=s|per=3|mod=i|ten=p	29	aux	_	_
29	stato	essere	V	VA	num=s|mod=p|gen=m	30	aux	_	_
30	trovato	trovare	V	V	num=s|mod=p|gen=m	26	mod_rel	_	_
31	nel	In	E	EA	num=s|gen=m	30	comp_temp	_	_
32	pomeriggio	pomeriggio	S	S	num=s|gen=m	31	prep	_	_
33	.	.	F	FS	_	0	punc	_	_

The full description of the adopted annotation schemes can be found at http://www.di.unito.it/~tutreeb/evalita-parsingtask-09.html (see in particular the section "Documents and Tools") for what concerns TUT and TUT-Penn, and at http://medialab.di.unipi.it/wiki/index.php/Tanl_Dependency_Tagset for TANL. It is worth pointing out here that, by carrying out both mainDepPar and pilotDepPar subtasks operating on different development sets, it will be possible to qualitatively and quantitatively evaluate the influence of the annotation features on parser performances. Starting from the comparison of the results of these two subtasks, it will be possible to bootstrap a larger unified resource for Italian to be hopefully used for the next EVALITA editions.

The blind version of the data for the test set, i.e. **TestSet-mainDepPar** and **TestSet-pilotDepPar**, will only contain morpho-syntactically annotated tokens, one per line; to be more concrete, the test data will contain the first six columns of the CoNLL format.

The format of the submitted run files must be the same as **DevSet-mainDepPar**, for the mainDepPar, and the same as **DevSet-pilotDepPar**, for the pilotDepPar, containing one token per line respectively with the corresponding TUT or TANL tags. In practice, parsed test data must be returned including all original columns of the test data plus the HEAD and DEPREL columns.

3.2.Constituency Parsing Data Format

The development corpus for Constituency Parsing is UTF-8 encoded, one sentence for each line as required by the EVALB evaluation metrics. The words are organized as in Penn Treebank format for what concerns the phrase structure of the sentence.

Example from TUT-Penn:

```
( (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)))))
    (. .)) )
```

In order to better describe the rich inflection of Italian, the PoS tagging has been instead developed especially for such a language; the full description of the PoS and functional tags adopted for TUT-Penn can be found at http://www.di.unito.it/~tutreeb/evalita-parsingtask-09.html (see in particular the section “Documents and Tools”).

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 **DevSetCosPar** with the corresponding TUT-Penn functional tags and phrase structure.

4. Evaluation Metrics

For both Dependency Parsing subtasks, the evaluation metrics are labeled and/or unlabeled attachment score: 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).

For the Constituency Parsing track, the evaluation metric is tree precision and recall no-crossing bracket metric calculated by using the EVALB program (Collins, 1996).

5. Evaluation Details

On January 15th 2009 the registration opens. For parsing task the participants will be requested to specify the tracks and subtasks of interest. The participation to both tracks (Dependency Parsing and Constituency Parsing) and/or subtasks of the Dependency Parsing (mainDepPar and pilotDepPar) is strongly encouraged.

On April 10th 2009 the task organizers will make available the development corpora on the Evalita web site for the registered participants (who have signed the license agreement).

On September 10th 2009 the organizers will make available the blind test set on the Evalita web site.

Participants should submit the results of their runs by September 20th 2009 (midnight, MDT), sending, to the organizers email address (`bosco@di.unito.it),` one file for each track and subtask they want to participate to, in the same format as the relative development corpus. Each file has to be named as:

`EVALITA09_PAR_SUB_ParticipantName`

where SUB has to be substituted by the name of the track (CosPar for Constituency Parsing) or subtask (mainDepPar for the obligatory main task of the Dependency Parsing track; pilotDepPar for the optional pilot task of the Dependency Parsing track).

Only one result file for each subtask and track will be accepted.

After the submission deadline, the organizers will evaluate the submitted runs and will send each participant the score of his submissions (October the 5th 2009) as well as the gold-standard version of the three 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] G. Attardi et al. (2008). TALN (Text Analytics and Natural Language processing). Project Analisi di Testi per il Semantic Web e il Question Answering, http://medialab.di.unipi.it/wiki/index.php/Analisi_di_testi_per_il_Semantic_Web_e_il_Question_Answering.

[8] S. Montemagni, M. Simi (2007). The Italian dependency annotated corpus developed for the CoNLL-2007 Shared Task. ILC Technical Report, January 2007, available at http://www.ilc.cnr.it/tressi_prg/ISST@CoNNL2007/ISST/ISST@CoNNL2007.pdf

[9] S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, R. Delmonte (2003), "Building the Italian Syntactic-Semantic Treebank", in Anne Abeillé (ed.), Building and using Parsed Corpora, Language and Speech series, Kluwer, Dordrecht, pp. 189-210.

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