The EVALITA Dependency Parsing Task: from 2007 to 2011

The EVALITA Dependency Parsing Task: From 2007 to 2011*

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 the EVALITA Parsing Task (EPT) is at defining and extending Italian state-of-the-art parsing by encouraging the application of existing models and approaches, comparing paradigms and annotation formats. Therefore, in all the editions, held respectively in 2007, 2009 and 2011, the Task has been organized around two tracks, namely Dependency Parsing and Constituency Parsing, exploiting the same data sets made available by the organizers in two different formats.

This paper describes the Dependency Parsing Task assuming an historical perspective, but mainly focussing on the last edition held in 2011. It presents and compares the resources exploited for development and testing, the participant systems and the results, showing also the improvement of resources and scores during the three editions of this contest.

Keywords: Dependency Parsing, Evaluation, Italian.

1 Introduction

The EVALITA Parsing Task (EPT) is an evaluation campaign which 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. In the current edition, held in 2011, as in the previous held respectively in 2007 [8,12] and 2009 [10,9], 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 the dependency track of the competition is analyzed mainly focussing on the more recent experience held in 2011, but also developing a comparison with the previous ones[1].

The paper is organized as follows. The next section describes the resource on which the EPT are based, i.e. the Turin University Treebank (TUT). Then we show a survey

* This work has been partially funded by the PARLI Project (Portale per l'Accesso alle Risorse Linguistiche per l'Italiano MIUR PRIN 2008).

[1] See [7] for a similar description of the constituency track.

B. Magnini et al. (Eds.): EVALITA 2012, LNCS 7689, pp. 1–12, 2012.

of the dependency parsing held in the 2007, 2009 and 2011 with all the information about data sets for training and testing, and the participation results. We conclude with a section where we compare and discuss the data presented in the other parts of the paper.

2 The Dataset for the EPT: The Turin University Treebank

TUT has been the reference resource for all the editions of the EPT, that is to say that the Parsing Task is based on the format of TUT and the data proposed for training and development of the participant systems were from this treebank for Italian. TUT is developed by the Natural Language Processing group of the Department of Computer Science of the University of Turin[2].

For each EPT edition, TUT has been ameliorated, by applying automatic and manual revisions oriented to improving consistency and correctness of the treebank, and enlarged, by adding new data also representing text genres new with respect to those attested in the resource. In particular, in 2011, TUT has been newly released and made as large as other existing Italian resources, i.e. Venice Italian Treebank (VIT, [21]) and ISST–TANL [17]. Moreover, in order to allow a variety of training and comparisons across various theoretical linguistic frameworks, during the last few years TUT has made available several annotation formats [4] beyond the native TUT, e.g. TUT–Penn, which is the conversion in a Penn–like format designed for Italian, and CCG–TUT, which is the conversion to the Combinatory Categorial Grammar for Italian [2].

2.1 The Native TUT Format

The native scheme of TUT applies the major principles of dependency grammar and exploits a rich set of grammatical relations [6,3]. In particular, among the existing dependency theoretical frameworks, TUT mainly follows the *Word Grammar* [15], and this is mirrored, for instance, in the annotation of determiners and prepositions as complementizers of nouns or verbs, and the selection of the main verb as head of the verbal structure instead of the auxiliary. For instance, in the example in table 2.1, the article *"La"* (The [Fem Sing]), and the prepositions *"a"* (in) play the head role respectively for the common noun *"coppia"* (couple) and the proper noun *"Milano"* (Milan); while the auxiliary verb *"stava"* (was [Progressive]) depends on the main verb *"trascorrendo"* (having).

For what concerns instead grammatical relations, which are the most typical feature of TUT, they are designed to represent a variety of linguistic information according to three different perspectives, i.e. morphology, functional syntax and semantics. The main idea is that a single layer, the one describing the relations between words, can represent linguistic knowledge that is proximate to semantics and underlies syntax and morphology, which seems to be unavoidable for efficient processing of human language, i.e. the predicate argument structure of events and states. Therefore, each relation label can in principle include three components, i.e. morpho-syntactic, functional-syntactic and

[2] For the free download of the resource, covered by a Creative Commons licence, see `http://www.di.unito.it/~tutreeb`

syntactic-semantic, but can be made more or less specialized, including from only one (i.e. the functional-syntactic) to three of them. For instance, among the relations used in the example in table 2.1, we can see that annotated on the node number 5 (corresponding to the lexical item *"a"* (in)); it represents the locative verbal indirect complement, i.e. VERB-INDCOMPL-LOC which includes all the three components and can be reduced to VERB-INDCOMPL (which includes only the first two components) or to INDCOMPL (which includes only the functional-syntactic component). This works as a means for the annotators to represent different layers of confidence in the annotation, but can also be applied to increase the comparability of TUT with other existing resources, by exploiting the amount of linguistic information more adequate for the comparison, e.g. in terms of number of relations, as happened in EPT (see below the TUT CoNLL format). Since in different settings several relations can be merged in a single one (e.g. VERB-INDCOMPL-LOC and INDCOMPL-LOC are merged in INDCOMPL), each setting includes a different number of relations: the setting based on the single functional-syntactic component includes 72 relations, the one based on morpho-syntactic and functional-syntactic components 140, and the one based on all the three components 323 [3,5].

Moreover TUT format 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. For instance, node 4.10 in table 2.1 represents the subject of the reduced relative clause headed by the verb *"residente"* (living); this subject, as usual in this kind of clause, is not lexically realized in the sentence, but TUT format, by using a null element and applying a co-indexing mechanism on it (with refers to the node number 2 corresponding to the lexical item *"coppia"* (couple)), allows the recovery of this subject. 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.

2.2 The TUT CoNLL Format

Nevertheless, in order to make possible the application of standard evaluation measures e.g. within EVALITA contests, the native format of TUT (see table 2.1) has been automatically converted in the standard CoNLL (see table 2) . 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, since only the functional syntactic component of the native TUT grammatical relations is taken into account); it does not include pointed indexes[3]. Since CoNLL does not allow null elements, they are deleted in this format,

[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 ....'.

Table 1. A sample sentence from the EPT 2009 test set as annotated in native TUT format: "La coppia, residente a Milano anche se di origini siciliane, stava trascorrendo un periodo di vacanza." (The couple, living in Milan even if of Sicilian provenance, was having a period of holiday.)

1	La	(IL ART DEF F SING)	[14;VERB-SUBJ]
2	coppia	(COPPIA NOUN COMMON F SING)	[1;DET+DEF-ARG]
3	,	(#, PUNCT)	[2;OPEN+PARENTHETICAL]
4	residente	(RISIEDERE VERB MAIN PARTICIPLE PAST INTRANS SING ALLVAL)	[2;VERB-RMOD+RELCL+REDUC]
4.10	t [2p]	(COPPIA NOUN COMMON F SING)	[4;VERB-SUBJ]
5	a	(A PREP MONO)	[4;VERB-INDCOMPL-LOC]
6	Milano	(MILANO NOUN PROPER F SING CITY)	[5;PREP-ARG]
7	anche	(ANCHE ADV CONCESS)	[8;ADVB+CONCESS-RMOD]
8	se	(SE CONJ SUBORD COND)	[4;VERB+FIN-RMOD]
8.10	t []	(ESSERE VERB MAIN IND PRES INTRANS 3 SING)	[8;CONJ-ARG]
9	di	(DI PREP MONO)	[8.10;VERB-PREDCOMPL+SUBJ]
10	origini	(ORIGINE NOUN COMMON F PL)	[9;PREP-ARG]
11	siciliane	(SICILIANO ADJ QUALIF F PL)	[10;ADJC+QUALIF-RMOD]
12	,	(#, PUNCT)	[2;CLOSE+PARENTHETICAL]
13	stava	(STARE VERB AUX IND IMPERF INTRANS 3 SING)	[14;AUX+PROGRESSIVE]
14	trascorrendo	(TRASCORRERE VERB MAIN GERUND PRES TRANS SING)	[0;TOP-VERB]
15	un	(UN ART INDEF M SING)	[14;VERB-OBJ]
16	periodo	(PERIODO NOUN COMMON M SING)	[15;DET+INDEF-ARG]
17	di	(DI PREP MONO)	[16;PREP-RMOD]
18	vacanza	(VACANZA NOUN COMMON F SING)	[17;PREP-ARG]
19	.	(#. PUNCT)	[14;END]

but the projectivity constraint is maintained at the cost of a loss of information with respect to native TUT in some cases. For instance, in the case of the ellipsis of a verbal head, the native TUT exploits a null element to represent it also linking the dependents to this null element, as in the usual case of the lexically realized verbal head; instead in the TUT CoNLL the verbal head remains missing and the dependents are linked where possible without violating the projectivity constraint.

3 The Dependency Parsing for Italian and the EVALITA Experience

As described in the Proceedings of the CoNLL Multilingual Shared Task [13,18], the Parsing Task is the activity of assigning a syntactic structure to a given set of Part of Speech 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 the evaluation of the official results of the Parsing Task, the metric exploited in the CoNLL contests is LAS (Labeled Attachment Score) that is the percentage of tokens with correct head and dependency type. Another measure often applied in parsing

Table 2. An example of annotation in TUT CoNLL format

1	La	IL	ART	ART	DEF|F |SING	14	SUBJ
2	coppia	COPPIA	NOUN	NOUN	COMMON|F|SING	1	ARG
3	,	#	PUNCT	PUNCT	_	2	OPEN+ PARENTHETICAL
4	residente	RISIEDERE	VERB	VERB	MAIN|PARTICIPLE|PAST| INTRANS|SING|ALLVAL	2	RMOD+RELCL +REDUC
5	a	A	PREP	PREP	MONO	4	INDCOMPL
6	Milano	MILANO	NOUN	NOUN	PROPER|F|SING|CITY	5	ARG
7	anche	ANCHE	ADV	ADV	CONCESS	8	RMOD
8	se	SE	CONJ	CONJ	SUBORD|COND	4	RMOD
9	di	DI	PREP	PREP	MONO	8	ARG
10	origini	ORIGINE	NOUN	NOUN	COMMON|F|PL	9	ARG
11	siciliane	SICILIANO	ADJ	ADJ	QUALIF|F|PL	10	RMOD
12	,	#	PUNCT	PUNCT	_	2	CLOSE+ PARENTHETICAL
13	stava	STARE	VERB	VERB	AUX|IND|IMPERF| INTRANS|3|SING	14	AUX+ PROGRESSIVE
14	trascorrendo	TRASCORRERE	VERB	VERB	MAIN|GERUND|PRES| TRANS|SING	0	TOP
15	un	UN	ART	ART	INDEF|M|SING	14	OBJ
16	periodo	PERIODO	NOUN	NOUN	COMMON|M|SING	15	ARG
17	di	DI	PREP	PREP	MONO	16	RMOD
18	vacanza	VACANZA	NOUN	NOUN	COMMON|F|SING	17	ARG
19	.	#	PUNCT	PUNCT	_	14	END

evaluation is UAS (Unlabeled Attachment Score), i.e. the percentage of tokens with correct head [13,18][4].

As far as the dependency parsing for Italian is concerned, the EVALITA evaluation campaigns adopted the same definition for the task and the same metric exploited and experienced within the CoNLL, i.e. LAS, but also UAS.

Also the most of information available before the EVALITA can be extracted from the Proceedings of the CoNLL Multilingual Shared Tasks, where Italian was among the analyzed languages. The best results published for Italian[5] are LAS 84.40, UAS 87.91, according to [18]. In particular, it should be noticed the performance of some parser which participated in the EPT too, i.e. DeSR, that achieved (81.34 LAS).

In the rest of this section, we describe the EPT held in 2007, 2009 and 2011 by showing the data sets exploited for training and testing the participant systems, and the results achieved by these parsers when applied on the test set.

3.1 EPT 2007

For the EPT 2007, the development set was composed by 2,000 sentences that correspond to 53,656 tokens[6] in the TUT CoNLL format. The organization of this set

[4] 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 [16].

[5] For English the reported results are LAS 88.11 and UAS 90.13 as in [19].

[6] Only words and punctuation marks are considered as tokens.

included two almost equally sized subcorpora including two different text genres, namely the Italian Civil Law Code (i.e. CODCIV, 25,424 tokens) and newspapers (i.e. NEWSPAPER, 28,232 tokens). The test set was instead composed by 200 sentences (4,962 tokens) and is balanced with respect to text genres as the training set.

Table 3. EPT 2007: results on the entire test set and on the two subcorpora (CODCIV and NEWSPAPER)

Participant	all testset		CODCIV		NEWSPAPER	
	LAS	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**
UniTo_Lesmo	86.94	90.90	92.37	93.59	81.50	88.21
UniPi_Attardi	77.88	88.43	82.47	92.06	71.23	85.02
IIIT_Mannem	75.12	85.81	76.33	88.76	73.91	82.86
UniStuttIMS_Schielen	74.85	85.88	77.18	89.95	72.51	81.80
UPenn_Champollion	*	85.46	*	88.30	*	82.61
UniRoma2_Zanzotto	47.62	62.11	48.14	64.86	47.09	59.36

Six different teams[7] participated in the task with the LAS and UAS scores reported in table 3. The average LAS calculated on the first four best scored systems[8] is 78.69, while the average UAS calculated on the same way is 87.75. Among the participant parsers, the UniTo_Lesmo parser, i.e. TULE (Turin University Linguistic Environment[9]), which resulted as the best scored, is featured by a rule-based approach, like UniRoma2_Zanzotto, while the others were statistical systems. TULE is a rule–based wide coverage parser developed in parallel with TUT by the Natural Language Processing group of the University of Turin, which has been applied to various domains. The second best scored is DeSR, a Shift/Reduce deterministic transition–based parser [1], which participated also in the CoNLL contests, as cited above.

As far as text genre is concerned, the best results refer to the data extracted from the CODCIV corpus. This result depends on the specific characteristics of the language exploited in legal texts, where more often than e.g. in newspaper texts the grammar rules are applied, but also on the structure of the Italian Civil Law Code, which includes

[7] The name of each system that participated in the contest is composed according to the following pattern: institution_author.

[8] Observing the amount of participants in the less participated edition of the EPT, i.e. that held in 2011, and in order to allow for comparison between the results in 2007, 2009 and 2011, we calculated the average taking into account only the four best scored participants. This is also motivated by the huge difference between the first five scored systems and the last one in both EPT 2007 and 2009, the inclusion of whose results in the average can be misleading; in fact the averages calculated on all the participants is very different: LAS is 72.48, while UAS is 83.09.

[9] `http://www.tule.di.unito.it/`

several very short sentences corresponding to the titles of articles or sections, which are obviously very easy to parse.

3.2 EPT 2009

For the EPT 2009, the training set included 2,400 sentences that correspond to 66,055 tokens in TUT CoNLL format. The corpus can be separated in three subcorpora, i.e. one from Italian newspapers (i.e. NEWSPAPER, 1,100 sentences and 30,561 tokens), one from the Italian Civil Law Code (i.e. CODCIV, 1,100 sentences and 28,048 tokens), and one from the Italian section of the JRC-Acquis Multilingual Parallel Corpus, a collection of declarations of the European Community[10] (200 sentences and 7,446 tokens). This small corpus (i.e. PASSAGE) includes text belonging to a new genre, and has been added in the data set for a collaboration between the EPT and the evaluation campaign for parsing French, Passage[11] that exploits texts from the corresponding French section of the same multilingual corpus.

The test set included 240 sentences (5,287 tokens) balanced as in the training set: 100 sentences (2,293 tokens) from Civil Law Code, 40 sentences (1,212 tokens) from the Passage/JRC-Acquis corpus, and 100 sentences (1,782 tokens) from newspapers. In particular, these latter sentences were included also in the test set of the pilot dependency parsing subtask organized for the EPT 2009 by the group of the Istituto di Linguistica Computazionale (ILC) and by the University of Pisa, see [10]. This subtask is based on another existing resource, the Italian Syntactic–Semantic Treebank (ISST–TANL, [17]) developed as a conjoint effort of the ILC and by the University of Pisa. It was mainly devoted to the development of comparisons between the formats respectively applied by TUT and ISST–TANL, as reported in [11]. It resulted in an assessment of the evaluation based on TUT, showing that quite close scores can be obtained also by exploiting the other Italian treebank.

Table 4. EPT 2009: results on the entire test set and on the subcorpora (CODCIV, NEWSPAPER and PASSAGE)

Participant	all testset		CODCIV		NEWSPAPER		PASSAGE	
	LAS	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**
UniTo_Lesmo	88.73	92.28	91.54	94.64	84.68	89.73	89.36	91.58
UniPi_Attardi	88.67	92.72	92.63	95.38	82.60	89.17	90.10	92.90
FBKirst_Lavelli	86.50	90.96	90.23	93.33	79.91	87.15	89.11	91.75
UniAmsterdam_Sangati	84.98	89.07	89.93	95.51	76.66	87.99	87.87	93.89
UniCopenhagen_Soegaard	80.42	89.05	86.04	90.27	72.84	81.93	80.94	85.31
CELI_Dini	68.00	77.95	70.74	74.97	63.86	70.15	68.89	73.35

[10] See `http://langtech.jrc.it/JRC-Acquis.html`
[11] See `http://atoll.inria.fr/passage/index.en.html`

The participants to the EPT 2009 were six and two were the best scored, since two parsers achieved results whose difference cannot be considered as statistically significant according to the p-value[12], namely UniTo_Lesmo and UniPi_Attardi. The former is an upgraded version of the rule-based parser that won the EPT in 2007, while the latter, i.e. DeSR, is the upgraded version of the second best scored in the same contest.

The best scores were again obtained on the data extracted from legal texts, while observing all the test set we see that it has been achieved 87.22 as average LAS and 91.25 as average LAS calculated on the four best scored[13].

3.3 EPT 2011

For the EPT 2011, the development set includes 3,452 Italian sentences (i.e. 94,722 in TUT CoNLL) and represents five different text genres organized in the following subcorpora:

- 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[14] (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)

The training set is therefore larger than before in particular with respect to the included text genres, i.e. WIKIPEDIA and COSTITA, which are newly included in the data set. As far as the test set is concerned, it is composed by 300 sentences (i.e. 7,836 tokens) around balanced as 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).

The participants to the dependency parsing track were four. Among them only one did not participate in the previous editions of the contest. Two participant systems, i.e. UniTo_Lesmo and Parsit_Grella, do not follow the statical approach. UniTo_Lesmo system is the rule–based parser, which won the EPTs in 2007 and 2009. The Parsit_Grella 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 is an application to Italian of different parsing algorithms implemented in MaltParser [19] and of an ensemble model made available by Mihai Surdeanu; UniPi_Attardi is instead DeSR, which participated in EPT in 2007 and won EPT in 2009.

According to the main evaluation measure, i.e. LAS, the best results have been achieved by Parsit_Grella followed by UniPi_Attardi (see table 5) with a difference statistically significant according to the p-value. The average scores of the participants are 88.76 for LAS and 93.55 for UAS. In table 5, we see also how the performance varies according to text genres. If evaluated on the civil law texts the difference among

[12] Note that the difference between two results is taken to be significant if $p < 0.05$, see `http://ilk.uvt.nl/conll/software.html`)

[13] We calculated the average as for the EPT 2007. The average LAS calculated on all participants is 82.88, while the average UAS is 87.96.

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

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 are significantly higher than those of the others, and this motivates the success of this parser in the contest.

Table 5. EPT 2011: results on the entire test set and on the subcorpora (CODCIV, NEWSPAPER and WIKIPEDIA)

Participant	all testset		CODCIV		NEWSPAPER		WIKIPEDIA	
	LAS	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**
Parsit_Grella	91.23	96.16	92.21	97.01	90.75	95.54	89.51	94.51
UniPi_Attardi	89.88	93.73	92.85	96.18	86.34	91.19	86.91	90.88
FBKirst_Lavelli	88.62	92.85	91.56	95.12	83.84	89.72	87.09	91.05
UniTo_Lesmo	85.34	91.47	89.06	94.43	80.69	87.70	81.87	88.80

4 Discussion

Observing the results showed in the paper for dependency parsing we can see an improvement from 2007 to 2011.

The best scores passed from 86.94 for LAS and 90.90 for UAS in 2007 (by UniTo_Lesmo), to 88.73–88.69 for LAS (by UniTo_Lesmo and UniPi_Attardi) and 92.72 for UAS (by UniPi_Attardi) in 2009, to 91.23 for LAS and 96.16 for UAS by Parsit_Grella in 2011. The average LAS is passed from 78.69 in 2007, to 87.22 in 2009, to 88.76 in 2011, while the average UAS from 87.75, to 91.25, to 93.55. The scores achieved in the last EPT positively compare also with the data for other languages, e.g. English (LAS $89,61\%$) and Japanese (LAS $91,65\%$) [18]. For what concerns text genres, in all the editions and tracks[15], the best performances are referred to the legal texts, while the other genres, namely Wikipedia and newspaper seem to be similarly harder to parse.

An analysis that goes beyond the mere scores should take into account various issues that can be related to this improvement.

First of all, even if it is very difficult to assess the amelioration of the quality of the data included in the EPT data sets, it is instead easy to see at least the increment in the size of the data sets exploited for training. As represented in figure 1, the data currently available for the development are almost the double of those available in 2007. There is a relationship between the improvement of results (see figure 1) and this increment of data sets. In particular, the larger amount of ameliorated available data has to be taken into account among the main motivations of the improvement of results for the statistical system which participated in all the EPTs from 2007 to 2011, i.e. DeSR.

[15] See the data about the constituency parsing track in [7].)

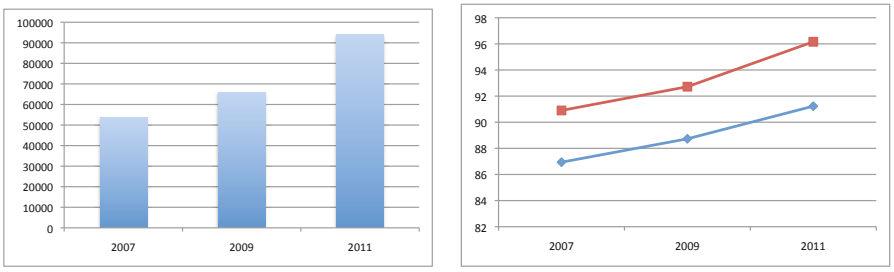

Fig. 1. The increment of the size of training data set and the improvement of best LAS scores during the three EPTs

We see also that in 2011, in contrast with EPT 2009 results, the top rule–based parser in 2009 and 2007 (UniTo_Lesmo) scores significantly worse than the two stochastic parsers (UniPi_Attardi and FBKirst_Lavelli). But the best performing system in 2011 is again a non pure system (i.e. Parsit_Grella). Nevertheless, also the results of this edition confirm that non–statistical systems can achieve good scores only if developed pursuing a continuous tuning on the reference resource, like UniTo_Lesmo in the past contests and Parsit_Grella today; while rule–based approaches not enough tuned on the resource obtained negative results, see e.g. [22] or [20].

Moreover, even if it is known in literature that it is very difficult to compare parsers that apply fundamentally different approaches, in order to allow for the participation in the EPT of both statical and rule-based approaches, the task has been always considered as open. This is to say that, since it is impossible to constrain the knowledge included in rule-based systems, also statistical parsers are admitted to be trained not only on the resources made available by the organizers of the EPT, but also on others in order to learn the knowledge needed for the application in the EPT. The exploitation of other sources of knowledge has been used, in particular, by the best scored parser of the last EPT, i.e. Parsit_Grella. This is a crucial issue to be taken into account for comparing the impressive results achieved by Parsit_Grella e.g. with those achieved by UniPi_Attardi, which follows instead a zero knowledge strategy learning all its knowledge only from the training data made available by the EPT organizers. More precisely, the exploitation of a lexicon and other linguistic data extracted from Wikipedia [14], explains the very good performance with respect to the other systems of Parsit_Grella on the WIKIPEDIA section of the test set.

The issues raised by the EVALITA experience in the Parsing Task are several and should be further investigated in the future. In particular, by assuming a wider perspective about the evaluation of the contribution of parsing to the overall quality of applicative NLP systems, we think that other kinds of information should be taken into account, e.g. those coming from null elements and semantic features currently annotated only in a few resources.

5 Conclusions

The EVALITA Parsing Tasks held during the last six years have been devoted to the definition and extension of the state-of-the-art for Italian parsing. Taking into account all the events of this evaluation campaign and mainly focussing on the last one held in 2011, the paper especially describes the evolution of the dependency parsing for Italian. It describes therefore the data sets used both in the training and evaluation, showing the details about the representation format implemented by TUT, namely the reference resource for the EPT experience. Then it describes the applied parsing systems and the results they achieved on the basis of these data in all the editions of the contest. Finally, a discussion about the results is presented.

References

1. Attardi, G., Simi, M.: DeSR at the Evalita Dependency Parsing Task. Intelligenza Artificiale 2(IV), 40–41 (2007)
2. Bos, J., Bosco, C., Mazzei, A.: Converting a dependency treebank to a Categorial Grammar treebank for Italian. In: Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT 2008), Milan, Italy, pp. 27–38 (2009)
3. Bosco, C.: A grammatical relation system for treebank annotation. Ph.D. thesis, University of Turin (2004)
4. Bosco, C.: Multiple-step treebank conversion: from dependency to Penn format. In: Proceedings of the Linguistic Annotation Workshop (LAW) 2007, Prague, Czech Republic, pp. 164–167 (2007)
5. Bosco, C., Lavelli, A.: Annotation schema oriented validation for dependency parsing evaluation. In: Proceedings of the Ninth Workshop on Treebanks and Linguistic Theories (TLT 2009), Tartu, Estonia, pp. 19–30 (2010)
6. Bosco, C., Lombardo, V., Vassallo, D., Lesmo, L.: Building a treebank for Italian: a data-driven annotation schema. In: Proceedings of second International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece (2000)
7. Bosco, C., Mazzei, A., Lavelli, A.: Looking back to the EVALITA Constituency Parsing Task: 2007-2011. In: Magnini, B., Cutugno, F., Falcone, M., Pianta, E. (eds.) EVALITA 2012. LNCS(LNAI), vol. 7689, pp. 46–57. Springer, Heidelberg (2012)
8. Bosco, C., Mazzei, A., Lombardo, V.: Evalita Parsing Task: an analysis of the first parsing system contest for Italian. Intelligenza Artificiale 2(IV), 30–33 (2007)
9. Bosco, C., Mazzei, A., Lombardo, V.: Evalita 2009 Parsing Task: constituency parsers and the Penn format for Italian. In: Proceedings of Evalita 2009, Reggio Emilia, Italy (2009)
10. Bosco, C., Montemagni, S., Mazzei, A., Lombardo, V., Dell'Orletta, F., Lenci, A.: Evalita 2009 Parsing Task: comparing dependency parsers and treebanks. In: Proceedings of Evalita 2009, Reggio Emilia, Italy (2009)
11. Bosco, C., Montemagni, S., Mazzei, A., Lombardo, V., Dell'Orletta, F., Lenci, A., Lesmo, L., Attardi, G., Simi, M., Lavelli, A., Hall, J., Nilsson, J., Nivre, J.: Comparing the influence of different treebank annotations on dependency parsing. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), La Valletta, Malta, pp. 1794–1801 (2010)
12. 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 the Sixth International Conference on Language Resources and Evaluation, Marrakech, Morocco, pp. 2066–2073 (2008)

13. Bucholz, S., Marsi, E.: CoNLL-X Shared Task on multilingual dependency parsing. In: Proceedings of the CoNLL-X, New York, USA, pp. 149–164 (2007)
14. Grella, M., Nicola, M., Christen, D.: Experiments with a constraint-based dependency parser. In: Evalita 2011 Working Notes (2012)
15. Hudson, R.: Word grammar. Basil Blackwell, Oxford (1984)
16. Kübler, S., McDonald, R., Nivre, J.: Dependency parsing. Morgan and Claypool Publishers (2009)
17. 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)
18. 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, Prague, Czech Republic, pp. 915–932 (2007)
19. Nivre, J., Hall, J.H., Chanev, A.: MaltParser: a language-independent system for data-driven dependency parsing. Natural Language Engineering 13(2), 95–135 (2007)
20. Testa, M., Bolioli, A., Dini, L., Mazzini, G.: Evaluation of a semantically oriented dependency grammar for Italian at EVALITA. In: Proceedings of Evalita 2009, Reggio Emilia, Italy (2009)
21. Tonelli, S., Delmonte, R., Bristot, A.: Enriching the Venice Italian Treebank with dependency and grammatical relations. In: Proceedings of the sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, pp. 1920–1924 (2008)
22. Zanzotto, F.M.: Lost in grammar translation. Intelligenza Artificiale 2(IV), 42–43 (2007)