EVALITA PARSING TASK: UN'ANALISI DELLA PRIMA COMPETIZIONE TRA SISTEMI DI ANALISI SINTATTICA PER LA LINGUA ITALIANA

EVALITA PARSING TASK: AN ANALYSIS OF THE FIRST PARSING SYSTEM CONTEST FOR ITALIAN

CRISTINA BOSCO · ALESSANDRO MAZZEI · VINCENZO LOMBARDO

SOMMARIO/*ABSTRACT*

EVALITA propone per la prima volta un confronto tra sistemi per il trattamento automatico della lingua italiana. L'articolo presenta, in particolare, un'analisi della competizione tra sistemi di parsing per l'italiano sia per il paradima a dipendenze che per quello a costituenti. Dopo una descrizione dei dati messi a disposizione dei partecipanti per lo sviluppo dei loro sistemi, si presenta una breve descrizione dei metodi standard applicati per valutare e confrontare i risultati prodotti dai sistemi in competizione. In conclusione alcune riflessioni e commenti sull'andamento della competizione.

EVALITA is the first attempt to analyze the NLP tools for Italian. This paper presents the results of the parsing competition on Italian by using dependency and constituency paradigms. We describe the data set provided to the participants of the task in order to develop their parsers and the methods used for evaluation. Finally we provide some conclusions on this parsing competition.

Keywords: Parsing, treebank, constituency, dependency, Italian.

1 Introduction

The application of parsing methods to different languages and corpora is currently considered a crucial and challenging task within the NLP international community. The validation of existing treebank-based parsing models, in fact, strongly depends on the possibility of generalizing their results on corpora and languages other than those on which they have been trained and tested.

In particular, for constituency-based parsing, several empirical evidences demonstrate the irreproducibility of the results obtained on the Penn Treebank on other treebanks, see e.g. [6], or languages, see e.g. [3] on Czech, [5] on German, [9] on Chinese, [4] on Italian. For dependency parsing, the results of the 2006 and 2007 CoNLL shared task showed that it is as robust as the constituency parsing, but equally affected by the problem of irreproducibility of results across corpora and languages [10, 11].

The aim of the EVALITA Parsing Task is at defining and extending the current state of the art in parsing Italian by encouraging the application of existing models, and contributing to the investigation of the causes of this irreproducibility. It provides the community members the possibility of focussing on Italian language and exploring different approaches. In fact, the development data are in various formats and the task is composed of subtasks with a quantitative evaluation of different kinds of outputs, also annotating different sets of features. Therefore the EVALITA Parsing Task is the first picture of the problems that lie ahead for Italian parsing and the kind of work necessary for adapting existing parsing models to this language.

The paper is organized as follows. The next section presents the data proposed to participant for the development and training of their systems, and the third section concerns, instead, the evaluation and results.

2 The development set

The data proposed for the development of parsing systems are from the Turin University Treebank (TUT) and are available at the web site `http://www.di.unito.it/~tutreeb`.

For dependency (TUT native format), the annotation applies the major tenets of dependency grammar [8] and implements an annotation schema which is based on a rich set of grammatical relations and centered upon the notion of argument structure. Moreover it includes null elements for the representation of non-projective structures, long distance dependencies, equi phenomena and pro drops, in order to allow for the representation and the recovery of the argument structure (associated with verbs and nouns).

For constituency (TUT-Penn format), the treebank adopts a Penn-like annotation which has been derived from the application of an automatic converter to the

dependency-based annotated data [2]. Like the Penn standard, TUT-Penn includes null elements (e.g. in relative clauses), but differentiates from Penn because of the PoS tagset[1].

Even if smaller than other existing Italian resources[2], TUT makes available more annotation formats that allow for a larger variety of training and testing for parsing systems. In fact the usefulness for the research community of a treebank is potentially limited by the degree to which a treebank subscribes to a specific linguistic theory, but the availability of several formats allows for comparatively testing the results of this task.

The development set includes 2.000 sentences that correspond to about 58.000 annotated tokens. The corpus can be separated in two equally sized subcorpora, one from the Italian legal Code and one from Italian newspaper. Both the subcorpora has been made available in various formats among which the following has been used by participant for tuning and training of parsing systems:

- the native TUT dependency format including null elements and featuring the TUT tokenization (i.e. with amalgams split in more lines with pointed indexes[3]), with each sentence identified by an index and followed by an empty line, and each line annotated according the pattern word_index/word/PoS/father_index/dep_relation
- the CoNLL standard format without null elements and featuring the TUT tokenization (i.e. with amalgams splitted in more lines without pointed indexes[4]) where the information included in the TUT native format has been splitted in 10 colums and standardized according to UTF-8 encoding and the CoNLL standard [10, 11]
- the Penn Treebank format that is the well-known constituency-based annotation revised for the application to Italian

3 The evaluation: test set, standard measures and results

The parsing task is defined as the activity of assigning a syntactic structure to a given Italian sentence using a fully automatic parser and according to the annotation schemes presented in the development set. In order to account for the large variety of parsing systems we have considered acceptable a number of discrepancies between the gold standard output (annotated according to TUT or TUT-Penn format as described before for the development set) and the participant output. Among these discrepancies we mention the absence of null elements both in dependency and constituency parsing, the absence (i.e. unlabeled dependency) or the underspecification of relation labels in dependency parsing (i.e. the annotation of the functional-syntactic component rather than of the three components of TUT relations). Other discrepancies have been tolerated and managed in order to allow for the evaluation of all the submitted results.

Among the 8 participants, 6 presented dependency parsing results, and two a constituency-based parses. Nobody tried both the tasks.

We have used two distinct procedures in order to evaluate either dependency and constituency parsing. For dependency we have used the three standard metrics used in the CoNLL parsing shared task: LAS (Labeled Attachment Score), i.e. the percentage of tokens with correct head and relation label; UAS (Unlabeled Attachment Score), i.e. the percentage of tokens with correct head; LAS2 (Label Accuracy) i.e. the percentage of tokens with correct relation label [10, 11]. For constituency we have used the standard brackets precision-recall-F_score metrics well known in parsing literature. As well as the development set, the test set was built on two different genres: one hundred sentences are from Italian legal Code and one hundred sentences are from Italian newspaper. The results of the parsers on the Italian legal Code test set are in Tables 3 and 5; on the Italian newspaper are in Tables 4 (dependency) and 6. The overall results are resumed in Tables 1 and 2.

The first clear result is that the Italian legal Code corpus is easier to parse than the Italian newspaper corpus. This is not really surprising since in general legal codes contains more regular sentences, but confirms some studies present in literature about the influence of the genre on parsing [12].

We can note that the best results for dependency format have been achieved by the UniTo_Lesmo_PAR parser. This rule-based parser has been developed in parallel with the TUT treebank, and so we can guess a certain influence over the annotators of the gold standard of the test set. The other parsers are statistics-based except UniRoma2_Zanzotto_PAR, again rule based. Statistics-based parsers have achieved notable results (although the development set is fifty times smaller than [11]), while the different tuning of the UniRoma2_Zanzotto_PAR rule-based parser can possibly explain the relatively poor performance.

For constituency format, the best result has been achieved by the UniNa_Corazza_PAR parser, again a statistical parser[5].

[1] The use of the Penn-TUT PoS tags, which are derived by reduction from the TUT original PoS tags, has been preferred to the Penn PoS tags since they better represent the inflectional richness of Italian.

[2] They are the Venice Italian Treebank (VIT) [7] and the Italian Syntactic Semantic Treebank (ISST) [1]

[3] E.g. the amalgamated word “del” in the 33th line of a sentence, is split in two lines, 33 del and 33.1 del respectively represent the Preposition and the Article.

[4] In CoNLL format each line is indexed by an index that corresponds exactly to the line number (within the single sentence).

[5] The errors reported in 2 depend on the different treatment by the parsers of the locutions with respect to the TUT-gold standard.

Table 1: Dependency parsing subtask evaluation

LAS	UAS	LAS2	Participant	Total
86.94	90.90	91.59	UniTo_Lesmo_PAR	1-1-1
77.88	88.43	83.00	UniPi_Attardi_PAR	2-2-2
75.12	85.81	82.05	IIIT_Mannem_PAR	3-4-3
74.85	85.88	81.59	UniStuttIMS_Schielen_PAR	4-3-4
*	85.46	*	UPenn_Champollion_PAR	*-5-*
47.62	62.11	54.90	UniRoma2_Zanzotto_PAR	5-6-5

Table 2: Constituency parsing subtask evaluation

Br-R	Br-P	Br-F	Errors	Participant
70.81	65.36	67.97	26	UniNa_Corazza_PAR
38.92	45.49	41.94	48	FBKirst_Pianta_PAR

Table 3: Dependency parsing subtask evaluation on legal Code subcorpus

LAS	UAS	LAS2	Participant	Total
92.37	93.59	95.86	UniTo_Lesmo_PAR	1-1-1
79.13	91.37	83.39	UniPi_Attardi_PAR	2-2-2
76.33	88.76	81.74	IIIT_Mannem_PAR	3-4-3
77.18	89.95	82.43	UniStuttIMS_Schielen_PAR	4-3-4
*	88.30	*	UPenn_Champollion_PAR	*-5-*
48.14	64.86	54.85	UniRoma2_Zanzotto_PAR	5-6-5

Table 4: Dependency parsing subtask evaluation on newspaper subcorpus

LAS	UAS	LAS2	Participant	Total
81.50	88.21	87.31	UniTo_Lesmo_PAR	1-1-1
76.62	85.49	82.61	UniPi_Attardi_PAR	2-2-2
73.91	82.86	82.35	IIIT_Mannem_PAR	3-4-3
72.51	81.80	80.74	UniStuttIMS_Schielen_PAR	4-3-4
*	82.61	*	UPenn_Champollion_PAR	*-5-*
47.09	59.36	54.94	UniRoma2_Zanzotto_PAR	5-6-5

Table 5: Constituency parsing subtask evaluation on legal Code subcorpus

Br-R	Br-P	Br-F	Errors	Participant
74.31	70.11	72.15	13	UniNa_Corazza_PAR
41.55	49.92	45.35	30	FBKirst_Pianta_PAR

Table 6: Constituency parsing subtask evaluation on newspaper subcorpus

Br-R	Br-P	Br-F	Errors	Participant
67.31	60.60	63.78	13	UniNa_Corazza_PAR
36.28	41.06	38.52	18	FBKirst_Pianta_PAR

4 Conclusions

The organization and the participation to the EVALITA parsing task have been big challenges for organizers as well as for participants. In order to compare systems it is necessary to adhere to standards, and this can be a not easy process.

Our impression is that for the dependency paradigm the parser involved in the competition are not so far from the state of art (i.e. parsers for English). In contrast, it seems that for constituency more effort is still necessary to achieve optimal results.

5 Acknowledgements

We thank for his support in the organizzation Fabio Zanzotto. Moreover we would like to thank Alberto Lavelli for his strong feedback on the deployment of the constituency data.

REFERENCES

[1] F. Barsotti and R. Basili and M. Battista and N. Calzolari and O. Corazzari and R. Del Monte and F. Fanciulli and N. Mana and M. Massetani and S. Montemagni and M.T. Pazienza and F. Pianesi and R. Raffaelli and D. Saracino and A. Zampolli and F.M. Zanzotto. *The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation* Kluwer, Dordrecht, Germany, 2001.

[2] C. Bosco. Multiple-step treebank conversion: from dependency to Penn format. *Proceedings of the Workshop on Linguistic Annotation at the ACL'07*, pp. 164–167, 2007.

[3] M. Collins and J. Hajič and L. Ramshaw and C. Tillmann. A Statistical Parser of Czech. *Proceedings of ACL 1999*, pp.505-512,1999.

[4] A. Corazza and A. Lavelli and G. Satta and R. Zanoli. Analyzing an Italian treebank with state-of-the-art statistical parser. *Proceedings of TLT-2004*, 2004.

[5] A. Dubey and F. Keller. Probabilistic parsing for German using sister-head dependencies. *Proceedings of the ACL'03*, 2003.

[6] D. Gildea. Corpus variation and parser performance. *Proceedings of the EMNLP'01*,2001.

[7] R. Delmonte. *Strutture sintattiche dall'analisi computazionale di corpora di italiano*, Franco Angeli, Milano, forthcoming.

[8] R. Hudson. *Word Grammar* Basil Blackwell, Oxford and New York, 1984.

[9] R. Levy and C. Manning. Is it harder to parse Chinese, or the Chinese treebank?. *Proceedings of the ACL'03*, 2003.

[10] S. Buchholz and and E. Marsi CoNLL-X Shared Task on Multilingual Dependency Parsing. *Proceedings of the CoNLL-X*, 2006.

[11] J. Nivre and J. Hall and S. Kübler and R. McDonald and J. Nilsson and S. Riedel and D. Yuret The CoNLL 2007 Shared Task on Dependency Parsing. *Proceedings of the EMNLP-CoNLL*, 2007.

[12] A. Mazzei and V. Lombardo A comparative Analysis of Extracted Grammars. *Proceedings of the ECAI*, 2004.

CONTACTS

CRISTINA BOSCO, ALESSANDRO MAZZEI, VINCENZO LOMBARDO
Dipartimento di Informatica, Università di Torino
Corso Svizzera 185 - 10149 Torino
Email: {bosco | mazzei | vincenzo}@di.unito.it

CRISTINA BOSCO is Assistant Professor at the Computer Science Department of the University of Torino where graduated in Philosophy, attended at a Master in Multimedia, earned her PhD in Computer Science and benefited of post-doc research grants. Her interests are syntax and linguistic resources. She is member of the ACL and participates at the TUT project.

ALESSANDRO MAZZEI is Assistant Professor at the Computer Science Department of the University of Torino. He graduated in Physics at the University Federico II of Naples (2000), and doctorated in Computer Science at the University of Turin (2005). His interests are in syntax, cognitive science and ontologies. He is member of the ACL.

VINCENZO LOMBARDO is Associate Professor at the Computer Science Department of the University of Torino. He graduated in Computer Science at the University of Turin (1987), and doctorated in Computer Science in the Turin-Milan University Consortium (1993). His interests are syntax, cognitive science and multimedia. He is member of the ACL.