The Berkeley Parser at the EVALITA 2011 Constituency Parsing Task*

Alberto Lavelli

FBK-irst,
via Sommarive 18, I-38123 Povo (TN), Italy
lavelli@fbk.eu

Abstract. In this paper we describe our participation at the EVALITA 2011 Constituency Parsing Task. We used the Berkeley Parser, obtaining an F_1 score of 83.83. This result corresponds to an increment of 6.48% with respect to the best result obtained at EVALITA 2009 by the Berkeley parser ($F_1 = 78.73$). This demonstrates that with treebanks of increasing size the results on Italian are improving and approaching those for other languages.

Keywords: Constituency parsing, statistical parsing, Italian.

1 Introduction

Our participation at the EVALITA 2011 constituency parsing task is part of a wider research effort devoted to the application of state-of-the-art statistical parsing techniques to Italian (see [7] for preliminary outcomes of such an effort). Statistical parsers can be ported to new languages by retraining them on a treebank for the new language. Quite often, they also require some knowledge about the new language, such as rules for the choice of heads in the grammar. We therefore compared the different tools not only on performance, but also regarding the manual interventions necessary for the porting.

Our participation to EVALITA started in 2007 [3], when we compared the Collins' parser [5], as implemented by Dan Bikel[1] [1], and the Stanford parser[2] [8, 9] (for more details on our participation at EVALITA 2007, see [6]). Adaptation of the parsers to Italian and to the Turin University Treebank (TUT) mainly consisted in the identification of rules for finding lexical heads. During the development phase of EVALITA 2007, we compared the performance of the Bikel's parser and of the Stanford parser, and the Bikel's parser was chosen for performing the official run on the 2007 test set.

The first row in Table 1 reports the results obtained in the official EVALITA 2007 evaluation. It should be noted that 26 sentences could not be evaluated, due to misalignment errors (i.e., sentences having a different number of words in the gold standard and in the parser output). Such errors were caused by the presence of multi-word expressions, which are usually taken into account during preprocessing. After the gold

* We thank Slav Petrov for making his parser available and for kindly answering our questions about its usage.

[1] http://www.cis.upenn.edu/~dbikel/#stat-parser

[2] http://nlp.stanford.edu/downloads/lex-parser.shtml

standard was released, we have run further experiments with multi-word expressions represented as single tokens. Such results are reported in the second and third rows of Table 1. In the fourth and fifth rows we show the results of the Bikel's parser in the leave-one-out (LOO) experiment on the development set, both considering all sentences and considering only sentences with less than 40 words.

Table 1. EVALITA 2007: Results on TUT using Parseval measures (LR: labeled recall; LP: labeled precision) and Exact Match Rate (EMR).

	LR	LP	F_1	EMR
EVALITA 2007 official results	70.81	63.35	67.96	
Bikel test	71.73	69.88	70.79	9.05
Bikel test < 40	72.04	70.08	71.05	9.84
Bikel LOO	73.42	72.49	72.95	18.43
Bikel LOO < 40	76.68	75.47	76.07	21.67

The results clearly confirm our previous experiments on ISST [7], with the Bikel's parser outperforming the Stanford parser. Therefore, in the 2009 edition we did not consider the Stanford parser, but we took into consideration a new parser, that is the Berkeley parser[3] [12], and compared its performance with the Bikel's parser, which obtained the best performance in the 2007 edition. As discussed above, in addition to performance we are also interested in the effort necessary to port the parser on a new language, that is Italian. The Berkeley parser seemed extremely interesting from this point of view as it requires no additional effort apart from the availability of a treebank.

The chosen experimental set-up was 10-fold cross validation (using LOO as in 2007 was not a viable option because of the time needed by the Berkeley parser to perform training). The training was performed using only the 19 basic PoS tags and the evaluation was done on the original treebank with full PoS tags (EVALB does not consider PoS accuracy when calculating Labeled Precision and Recall) and without considering punctuation.

Table 2. EVALITA 2009: Parser trained using basic PoS tags and evaluated on the original treebank with full PoS tags. Results obtained using 10-fold cross validation on the training set.

	LR	LP	F_1	EMR
Bikel	71.65	70.89	71.27	15.61
Bikel < 40	75.18	74.12	74.64	18.89
Berkeley - iteration #4	78.51	79.47	78.99	25.93
Berkeley - iteration #4 < 40	81.75	82.45	82.10	31.35

The Berkeley parser with the grammar obtained at iteration #4 was chosen for performing the official run on the 2009 test set. In [2] the official results for the constituency

[3] `http://nlp.cs.berkeley.edu/Main.html#Parsing`

parsing task can be found. Our system obtained the best result (F_1: 78.73; R: 80.02; P: 77.48). In Table 3 the results obtained on the test set by the Bikel's parser and by the Berkeley parser with grammars at different iterations are shown. The results confirm that the Berkeley parser outperforms the Bikel's parser. Note that in the official evaluation punctuation was taken into account and this greatly affected the performance of Bikel's parser (see Table 4 for results on the test set without considering punctuation).

The results show that the Berkeley parser performs better than the Bikel's parser and moreover do not require any language-specific adaptation. This is in line with what reported in the [16] where different parsers are compared on French and the Berkeley parser wins over the other parsers.

Table 3. EVALITA 2009: Results obtained by the Bikel's parser and by the Berkeley parser on the test set.

	LR	LP	F_1	EMR
Bikel	68.51	64.45	66.42	14.00
Bikel < 40	68.99	65.03	66.95	14.81
Berkeley - iteration #4	80.02	77.48	78.73	21.00
Berkeley - iteration #4 < 40	79.90	77.92	78.90	22.22

Table 4. EVALITA 2009: Results obtained by the Bikel's parser and by the Berkeley parser on the test set without considering punctuation.

	LR	LP	F_1	EMR
Bikel	74.08	69.70	71.82	14.00
Bikel < 40	74.74	70.45	72.53	14.81
Berkeley - iteration #4	80.20	77.65	78.90	21.00
Berkeley - iteration #4 < 40	80.11	78.12	79.10	22.22

2 Participation at EVALITA 2011

For the 2011 EVALITA edition we applied again the Berkeley parser[4] [12]. As discussed above, in addition to performance we are also interested in the effort necessary to port the parser on a new language, that is Italian. Berkeley parser seemed extremely interesting from this point of view as it requires no additional effort apart from the availability of a treebank.

The Berkeley parser is based on a hierarchical coarse-to-fine parsing, where a sequence of grammars is considered, each being the refinement, namely a partial splitting, of the preceding one. Its performance is at the state of the art for English on the Penn

[4] `http://nlp.cs.berkeley.edu/Main.html#Parsing`

Treebank and it outperforms other parsers in languages different from English, namely German and Chinese [12]. Indeed, a good compromise between efficiency and accuracy is obtained by a node splitting procedure, where splits which do not help accuracy are immediately pruned. Training is based on a discriminative framework, as discussed in [13]. As we aim at maximizing F_1, we used the parser version without reranking according to likelihood.

3 Results at EVALITA 2011

In 2011 we used the Berkeley parser only. Our initial plans were to use a reranking approach (adapting the software made available by David McClosky) but we did not manage to do it.

3.1 EVALITA 2011 dataset

The TUT version used in 2011 as training set consisted of 3,542 sentences, 1,983 from legal texts, 1,100 from newspapers and 459 from Wikipedia. The test set was composed by 300 sentences (150 from legal texts, 75 from newspapers and 75 from Wikipedia). The PoS tag set consists of 19 basic tags (68 including morphological features) and 29 nonterminal symbols.

As we did in previous editions, we specialized the PUNCT PoS tag associated to punctuation to more specific PoS tags, similarly to what is done in the PennTreeBank annotation.

3.2 Experimental Results on the 2011 Training Set

First of all, we report the results obtained on the training set. The chosen experimental set-up was 10-fold cross validation. As in 2009 this produced the best results for the Berkeley parser, in our experiments we used as PoS tagset only the 19 basic PoS tags. The rationale of such setting was to reduce data sparsity. In Table 5 the results obtained performing the training of the parser using basic PoS tags are displayed. The evaluation was done on the original treebank with full PoS tags (EVALB does not consider PoS accuracy when calculating Labeled Precision and Recall) and without considering punctuation.

Table 5. EVALITA 2011: Parser trained using basic PoS tags and evaluated on the original treebank with full PoS tags. Results obtained using 10-fold cross validation on the training set.

	LR	LP	F_1	EMR
Berkeley	78.74	79.32	78.99	26.33
Berkeley < 40	81.88	82.38	82.10	31.82

3.3 Experimental Results on the 2011 Test Set

The Berkeley parser performed the official run on the 2009 test set. In Table 3 the results obtained on the test set by the Berkeley parser are shown.

Table 6. EVALITA 2011: Results obtained by the Berkeley parser on the test set.

	LR	LP	F_1	EMR
Berkeley	83.54	84.12	83.83	22.74
Berkeley < 40	83.69	84.35	84.02	24.20

4 Conclusions

In this paper we have described our participation at the EVALITA 2011 Constituency Parsing Task. We used the Berkeley Parser, obtaining an F_1 score of 83.83. This result corresponds to an increment of 6.48% with respect to the best result obtained at EVALITA 2009 by the Berkeley parser ($F_1 = 78.73$). This demonstrates that with treebanks of increasing size the results on Italian are improving and approaching those for other languages.

We plan to keep working on improving parsing results on Italian, experimenting e.g. the Charniak reranking parser [4] and the use of self training both with reranking [10, 11] and without reranking [15]. Moreover, we are currently integrating the parser in the TextPro tool suite [14] to make it usable within other more complex systems (e.g., textual entailment, question answering, ...).

References

1. Bikel, D.M.: Intricacies of Collins' parsing model. Computational Linguistics 30(4), 479–511 (2004)
2. Bosco, C., Mazzei, A., Lombardo, V.: Evalita'09 Parsing Task: constituency parsers and the Penn format for Italian. In: Proceedings of the EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian (2009)
3. 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 (May 2008), http://www.lrec-conf.org/proceedings/lrec2008/pdf/528_paper.pdf
4. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and maxent discriminative reranking. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). pp. 173–180. Ann Arbor, Michigan (June 2005), http://www.aclweb.org/anthology/P05-1022
5. Collins, M.: Head-Driven Statistical Models for Natural Language Parsing. Ph.D. thesis, University of Pennsylvania (1999)

6. Corazza, A., Lavelli, A., Satta, G.: Phrase-based statistical parsing. In: Proceedings of the EVALITA 2007 Workshop on Evaluation of NLP Tools for Italian (2007)
7. Corazza, A., Lavelli, A., Satta, G., Zanoli, R.: Analyzing an Italian treebank with state-of-the-art statistical parsers. In: Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004). Tübingen, Germany (2004)
8. Klein, D., Manning, C.D.: Fast exact inference with a factored model for natural language parsing. In: Advances in Neural Information Processing Systems 15 (NIPS 2002) (2002)
9. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan (2003)
10. McClosky, D., Charniak, E., Johnson, M.: Effective self-training for parsing. In: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. pp. 152–159. New York City, USA (June 2006), http://www.aclweb.org/anthology/N/N06/N06-1020
11. McClosky, D., Charniak, E., Johnson, M.: Reranking and self-training for parser adaptation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. pp. 337–344. Sydney, Australia (July 2006), http://www.aclweb.org/anthology/P06-1043
12. Petrov, S., Klein, D.: Improved inference for unlexicalized parsing. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. pp. 404–411. Rochester, New York (April 2007), http://www.aclweb.org/anthology/N/N07/N07-1051
13. Petrov, S., Klein, D.: Discriminative log-linear grammars with latent variables. In: Proceedings of NIPS 20 (2008)
14. Pianta, E., Girardi, C., Zanoli, R.: The TextPro tool suite. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation. Marrakech, Morocco (May 2008), http://www.lrec-conf.org/proceedings/lrec2008/pdf/645_paper.pdf
15. Reichart, R., Rappoport, A.: Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 616–623. Prague, Czech Republic (June 2007), http://www.aclweb.org/anthology/P07-1078
16. Seddah, D., Candito, M., Crabbé, B.: Cross parsers evaluation: a French treebanks study. In: Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09). Paris, France (October 2009)