A simple DOP model for constituency parsing of Italian sentences

Federico Sangati

Institute for Logic, Language and Computation - University of Amsterdam
f.sangati@uva.nl

Abstract. We present a simplified Data-Oriented Parsing (DOP) formalism for learning the constituency structure of Italian sentences. In our approach we try to simplify the original DOP methodology by constraining the number and type of fragments we extract from the training corpus. We provide some examples of the types of constructions that occur more often in the treebank, and quantify the performance of our grammar on the constituency parsing task.

Keywords: Data-Oriented Parsing, Tree substitution grammar, statistical model, fragments, kernel methods.

1 Introduction

The Data-Oriented Parsing (DOP) framework, proposed in [1] and developed in [2], has become one of the most successful methods in constituency parsing (cf. [3], [4]). The main idea behind this methodology is to extract as many as possible fragments from the training corpus, and recombine them via a probabilistic generative model, in order to parse novel sentences. In the current EVALITA'09 task we aim at simplifying the original DOP methodology by constraining the number of fragments we extract from the training corpus. In particular we maintain only those fragments which are occurring at least two times in the training data. The main motivation behind this choice is to keep in our grammar only those fragments for which there is an empirical evidence about their reusability.

1.1 Data-Oriented Parsing

A DOP grammar can be described as a collection T of *fragments*. Figure 1 shows an example of four fragments that are extracted from the training parse tree depicted in figure 2, belonging to the TUT[1] training corpus. Fragments are defined in such a way that every node is either a non-terminal leaf (with no more daughters), or has the exact same daughters as in the original tree.

Two elementary trees α and β can be combined by means of the *substitution operation*, $\alpha \circ \beta$, iff the root of β has the same label of the leftmost nonterminal leaf of α. The result of this operation is a unified fragment which corresponds to α with the leftmost nonterminal leaf replaced with the entire fragment β. The substitution operation can be applied iteratively: $\alpha \circ \beta \circ \gamma = (\alpha \circ \beta) \circ \gamma$.

[1] Turin University Treebank: `http://www.di.unito.it/~tutreeb` , see also [5].

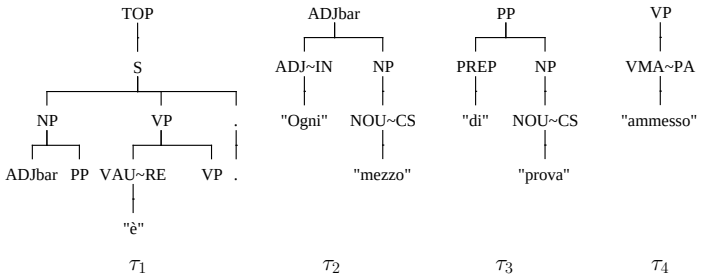

Fig. 1: Example of elementary trees of depth 4, 3, 3, and 2.

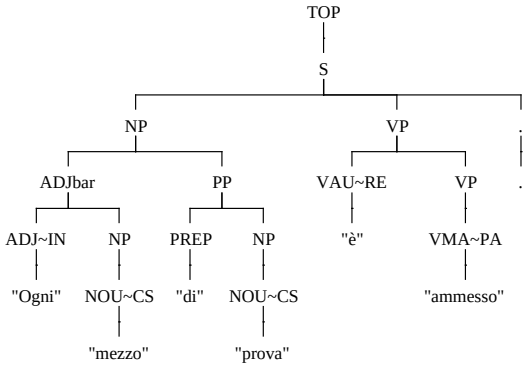

Fig. 2: Parse tree of the sentence *"Ogni mezzo di prova è ammesso"* (Every piece of evidence is admitted).

When the tree resulting from a series of substitution operations is a complete parse tree, i.e. all its leaf nodes are lexical nodes, we define the sequence of the elementary trees used in the operations as a *derivation* of the complete parse tree. Considering the 4 elementary trees in figure 1, $\tau_1 \circ \tau_2 \circ \tau_3 \circ \tau_4$ constitutes a possible derivation of the complete parse tree of Figure 2.

A stochastic instantiation of this grammar can be defined as follow: for every $\tau \in T$, the probability of using τ in a substitution operation is defined as $P(\tau) = \frac{f(\tau,T)}{f(root(\tau),T)}$, where the numerator returns the frequency of τ in T, and the denominator the number of fragments in T having $root(\tau)$ as root node. If a derivation d is constituted by n elementary trees $\tau_1 \circ \tau_2 \circ \ldots \circ \tau_n$, the probability of the derivation is calculated as: $P(d) = \prod_{i=1}^{n} P(\tau_i)$. Given that we have multiple derivations $d_1, d_2, \ldots, d_m$ for the same parse tree t, the probability of t is defined as: $P(t) = \sum_{i=1}^{m} P(d_i)$.

2 Implementation

In order to build our DOP grammar we have extracted all the fragments occurring in the 2,200 training structures[2] two or more times, by using an algorithm which is analogous to the one presented in [6]. In figure 3 we show the distribution of the frequencies of the extracted fragments with respect to their depths. In figure 4 we report the most common fragments containing the verb *è* (is), which can be seen as a collection of its main valency structures appearing in the annotated data. In addition to these fragments we have added in our grammar all CFG rules that occur exactly once in the training corpus (9,497 rules).

We have converted the DOP grammar to an isomorphic CFG (more details in [7]), and used the BitPar parser in [8] to parse the 200 sentences in the test set. For every test sentence we have approximated the most probable parse tree by taking the 1,000 most probable derivations, summing the probabilities of those yielding the same parse tree, and selecting the most probable.

3 Results

Table 1 shows a summary of the parsing results of our system, which achieves 75.76% in labeled F-score. More detailed analyses on the results are given in figure 6 where we show the accuracy within each single label: all main categories (NP, PP, VP, S) achieve an accuracy which is in line with the overall score of the system. Further investigations presented in figure 5 and in figure 7 suggest that the majority of parsing errors are due to crossing brackets among these four categories; wrongly labeled constituents are in fact a minor source of error.

4 Conclusions

We have presented a simplified DOP formalism for learning the constituency structure of Italian sentences. As in previous works (cf. [7], [9], [10]) the main motivation was to try to build a grammar based on those structures which are linguistically relevant, in this case those for which there is some empirical evidence about their reusability. The results are poor relative to the same methodology applied to English treebanks. One of the main reasons is certainly the smaller size of the training corpus used in the current shared task: as in other types of exemplar-based learning techniques, DOP models require a large amount of data in order to achieve high accuracy. We nevertheless believe that few more steps could contribute to improve results within the same framework, in particular the use of proper smoothing techniques over the fragments as in [11], and an investigation over different probability distributions.

Acknowledgments We gratefully acknowledge funding by the Netherlands Organization for Scientific Research (NWO): the author is funded through a Vici-grant "Integrating Cognition" (277.70.006) to Rens Bod.

[2] We have removed all empty nodes, traces, and functional labels from the corpus.

Depth	Types	Tokens
1	3364	88074
2	5818	72718
3	8768	60651
4	8328	37745
5	4795	16047
6	1839	5035
7	560	1343
8	118	248
9	29	61
10	6	13
11	2	4
13	1	2
14	1	2

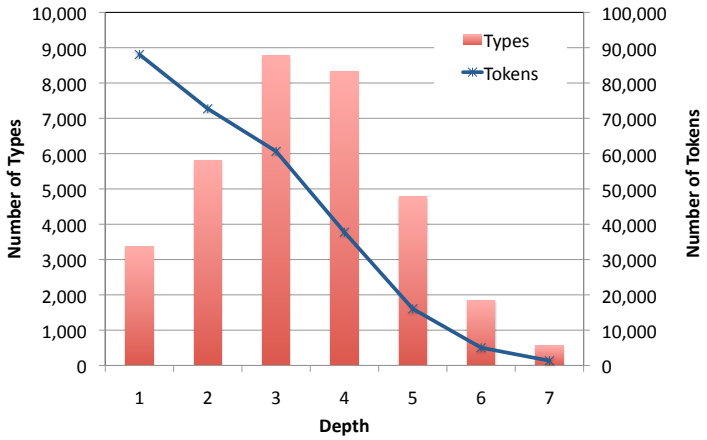

Fig. 3: Distribution of the frequency of the extracted 33,629 fragments with respect to their depths. All these fragments occur at least two times in the training corpus.

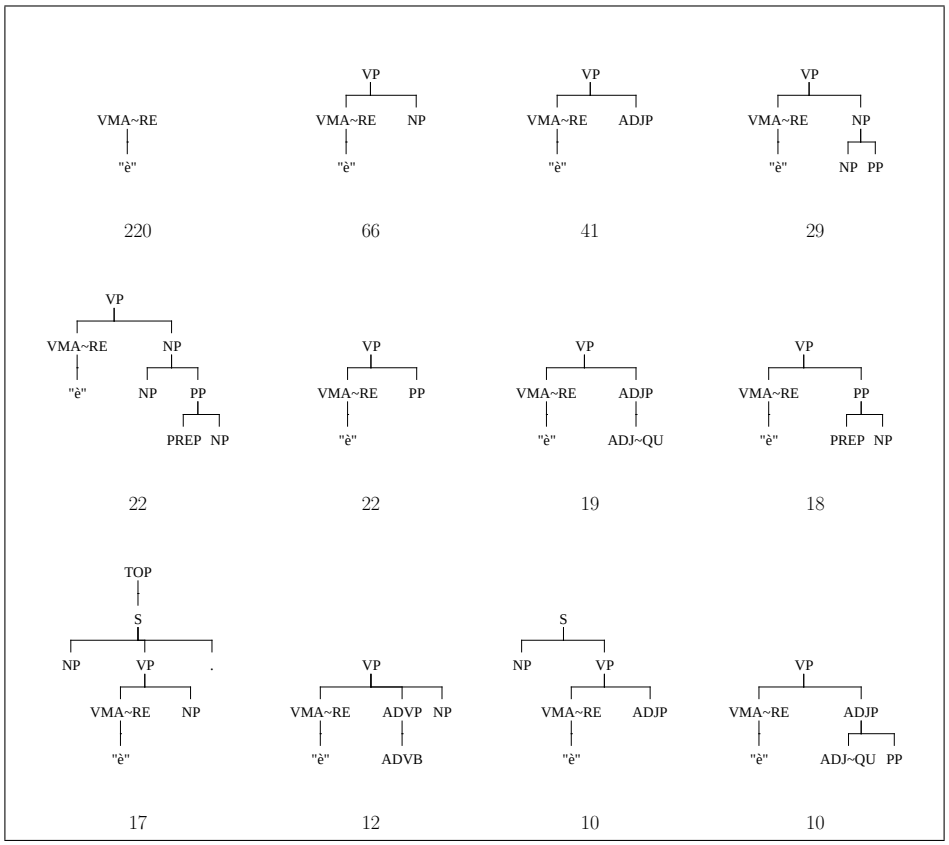

Fig. 4: The most frequent fragments in the grammar containing the verb *è* (*is*), when it is a main verb (VMA∼RE) and not an auxiliary (VAU∼RE).

Table 1: Summary of the parsing evaluation results.

Bracketing Labeled Recall	78.53%
Bracketing Labeled Precision	73.24%
Bracketing Labeled F-score	75.79%
Complete match	20.00%
Average crossing	2.47
No crossing	42.50%
2 or less crossing	66.00%

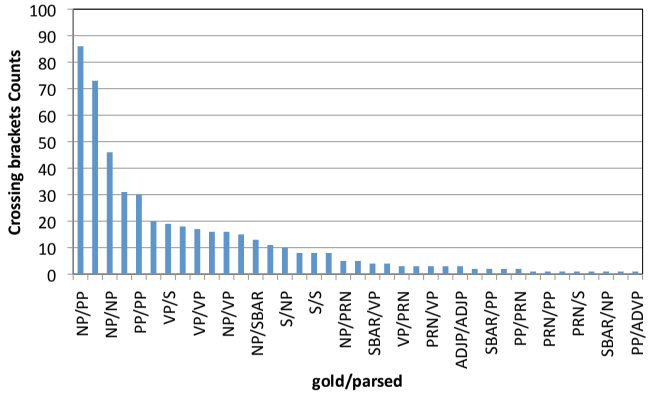

Fig. 5: Frequencies of crossing brackets (number of constituents in the gold tree that cross one constituent in the parsed tree.)

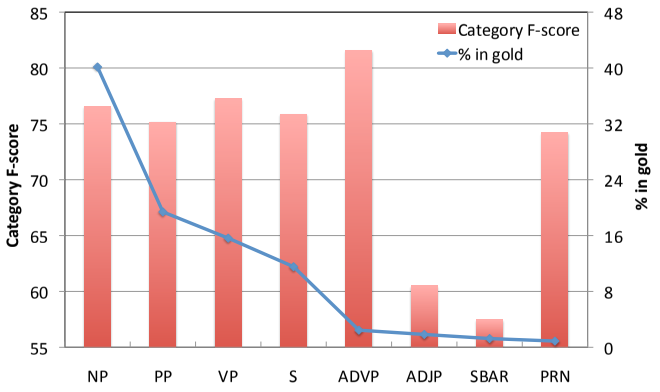

Fig. 6: F-score of the 8 most frequent categories in the corpus.

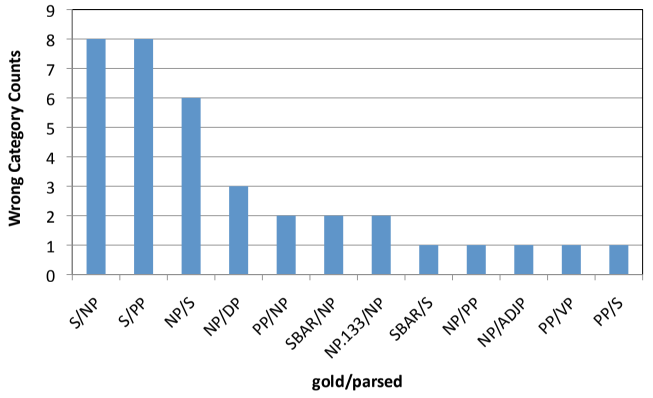

Fig. 7: Frequencies of constituents in the parsed tree with correct spans but wrong labels.

References

1. Scha, R.: Taaltheorie en taaltechnologie; competence en performance. In: de Kort, R., Leerdam, G., (eds.) Computertoepassingen in de Neerlandistiek, pp. 7–22. LVVN, Almere, the Netherlands. English translation at http://iaaa.nl/rs/LeerdamE.html (1990)
2. Bod, R.: A computational model of language performance: Data oriented parsing. In: Proceedings of COLING 1992, pp. 855–859 (1992)
3. Bod, R.: What is the minimal set of fragments that achieves maximal parse accuracy? In: Proceedings of ACL 2001, pp. 66–73 (2001)
4. Bod, R., Sima'an, K., Scha, R.: Data-Oriented Parsing. University of Chicago Press, Chicago, IL, USA (2003)
5. Lesmo, L., Lombardo, V., Bosco, C.: Treebank development: the TUT approach. In: Proceedings of the International Conference on Natural Language Processing, pp. 61–70. Vikas Publishing House (2002)
6. Collins, M., Duffy, N.: Convolution kernels for natural language. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) NIPS, pp. 625–632. MIT Press (2001)
7. Sangati, F.: Towards simpler tree substitution grammars, MSc Thesis (2007)
8. Schmid, H.: Efficient parsing of highly ambiguous context-free grammars with bit vectors. In: Proceedings of COLING 2004, pp. 162–168. Geneva, Switzerland (2004)
9. Sangati, F., Zuidema, W.: Unsupervised methods for head assignments. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pp. 701–709. Association for Computational Linguistics (2009)
10. Zuidema, W.:Parsimonious data-oriented parsing. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 551–560. Association for Computational Linguistics (2007)
11. Bod, R.: Two questions about data-oriented parsing. In: Proceedings of the Fourth Workshop on Very Large Corpora (1996)