Parsing with Intraclausal Coordination and Clause Detection

Domen Marinčič
Institut “Jožef Stefan”, Jamova cesta 39, 1000 Ljubljana
E-mail: domen.marincic@ijs.si

Thesis Summary

Keywords: clause and intraclausal coordination detection, parsing, machine learning

Received: January 26, 2010

This paper presents the work on syntactic analysis of Slovene text. A new algorithm for parsing using intraclausal coordination and clause detection is described. The experiments show that the algorithm achieves a significant decrease in the number of parsing errors.

Povzetek: Članek opisuje nov algoritem za skladenjsko razčlenjevanje z iskanjem naštevanj in stavkov.

1 Introduction

Syntactic analysis, i.e., parsing of text is used during various tasks, e.g., machine translation, question answering, etc. The structure of a sentence is represented with a tree. Parsing long sentences is a difficult task. The motivation was to analyze sub-units of the sentence independently, which could improve the overall parsing accuracy. We developed a new parsing algorithm that includes intraclausal coordination and clause detection.

Parsing using clause detection was first tried by Abney (1), whose algorithm delimits non-embedded clauses before the complete parse is made. In (2), there is a short description of a rule-based parser where clause identification is included in the parsing process. A detailed description of our new algorithm can be found in (3).

To our knowledge, the algorithm is the first to use intraclausal coordination detection in cojunction with clause detection before parsing. The most important contribution is the decrease in the number of parsing errors by 7.1% and 6.4% for Slovene, compared to the Malt (4) and MSTP (5) baseline parsers, respectively.

2 The algorithm

The first phase is a loop for intraclausal coordination and clause detection. It begins by splitting the sentence into segments. Punctuation tokens and conjunctions are delimiters between the segments (the vertical line in Fig. 1). Then, the intraclausal coordinations are detected and reduced into the meta tokens. In the example in Fig. 1, one intraclausal coordination is found. In the next step, the sentence is split into segments again. At the end, clause detection and reduction is performed. The loop iterates until no more units can be retrieved or only one segment remains. Detection in the example sentence in Fig. 1 finishes in the step b).

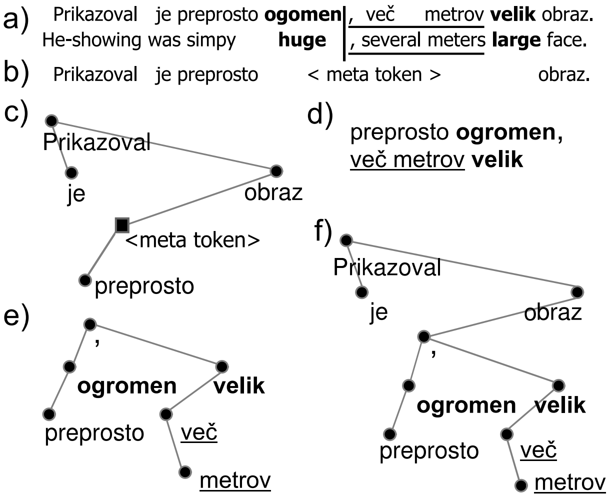
a) Prikazoval je preprosto ogomen , več metrov velik obraz.
He-showing was simpy huge , several meters large face.
b) Prikazoval je preprosto < meta token > obraz.
c)
Prikazoval
je
obraz
<meta token>
preprosto
d)
preprosto ogromen,
več metrov velik
e)
,
ogromen
velik
preprosto
več
metrov
f)
Prikazoval
je
obraz
,
ogromen
velik
preprosto
več
metrov

Figure 1: An example how a sentence is processed.

Detection of intraclausal coordinations and clauses is made in two steps: (i) candidate search and (ii) candidate classification using the AdaboostM1 algorithm. The candidates for intraclausal coordinations are searched for with the heuristic rule, stating that all the head words (in bold in Fig. 1a) must have the same part-of-speech and case. The candidates are then machinely classified using the features (presence of an adverb, noun/adjective matching with the head word) from the text between the head words, underlined in Fig. 1a. For the clause candidates, all the verb segments are taken. The following features present in the segment are used for machine classification: conjunctions, pronouns, punctuation tokens, auxiliary verbs, possible crossing intraclausal coordinations. The positively classified candidates are reduced.

The second phase builds the parse tree. It begins by parsing the sequence remaining after the first phase by the base parsers into the initial sentence tree, Fig. 1c. Certain errors in the initial tree are corrected by a newly developed rule-based parser. Then, the meta tokens are processed in a loop containing three steps: (i) the tokens of the meta-token subtree are joined with the unit that corresponds with the meta token, Fig. 1d; (ii) the new sequence is parsed, Fig. 1e; (iii) the subtree is merged with the sentence tree, Fig. 1f.

3 Evaluation

The experiments for estimating parsing accuracy are presented (Table 1). The part of the SDT corpus (6) from the Orwell's novel "1984" was used as the train and test set. Each experiment was carried out either with the MSTP parser or the Malt parser in the role of the base parsers. As the accuracy measure, the quotient between the nodes (punctuation excluded) assigned the correct parent and all the nodes in the tree was used. The accuracies of the plain MSTP and Malt parsers represent the baseline results.

Various versions of the new algorithm were compared: (i) the baseline parsers without detection; (ii) detection without classification and the rule-based parser; (iii) the classifiers turned on, no rule-based parser; (iv) the full version, achieving the 6.4% and 7.1% relative decrease of error compared to the baseline results.

Parsing algorithm	Malt	MSTP
Baseline	73,28 %	80,24 %
No classif., no rule-based p.	*74,63 %	*81,05 %
No rule-based parser	*74,83 %	*81,34 %
Full detection	*75,19 %	*81,51 %

Table 1: Parsing accuracy. The results marked with * are statistically significantly different from the baseline results.

4 Conclusion

The experiments show that by dividing complex sentences into smaller, more easily manageable units parsing accuracy can be increased. This was made possible by encoding background knowledge about the structure of clauses and intraclausal coordinations into the heuristic rules and classifiers used at the detection phase. Such knowledge apparently cannot be mined from the data by language-independent parsers. The most important idea for the future work seems to be the following: encode the information about the intraclausal coordination and clause structure as additional features of the words to enable a parser to combine this information with other knowledge about the parsed text more smoothly.

References

[1] Abney, S. P. (1990) In Proc. of the 6th New OED Conference : pp. 1–9.

[2] Holán, T. and Žabokrtský, Z. (2006) In Proc. of the TSD conference : pp. 95–102.

[3] Marinčič, D. Automatic text parsing with intraclausal coordination and clause detection, PhD thesis , Jozef Stefan International Postgraduate School (2008).

[4] Nivre, J. (2006) Inductive Dependency Parsing, Springer, The Netherlands.

[5] McDonald, R., Pereira, F., Ribarov, K., and Hajié, J. (2005) In Proc. of the HLT-EMNLP conference : pp. 523–530.

[6] Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtský, Z., and Žele, A. (2006) In Proc. of the LREC conference : pp. 1388–1391.