Best Analysis Selection in Inflectional Languages

Aleš Horák and **Pavel Smrž**
Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: {hales,smrz}@fi.muni.cz

Abstract

Ambiguity is the fundamental property of natural language. Perhaps, the most burdensome case of ambiguity manifests itself on the syntactic level of analysis. In order to face up to the high number of obtained derivation trees, this paper describes several techniques for evaluation of the figures of merit, which define a sort order on parsing trees. The presented methods are based on language specific features of synthetical languages and they improve the results of simple stochastic approaches.

1 Introduction

Ambiguity on all levels of representation is an inherent property of natural languages and it also forms a central problem of natural language parsing. A consequence of the natural language ambiguity is a high number of possible outputs of a parser that are usually represented by labeled trees. The average number of parsing trees per input sentence strongly depends on the background grammar and thence on the language. There are natural language grammars producing at most hundreds or thousands of parsing trees but also highly ambiguous grammar systems producing enormous number of results. For example, a grammar extracted from the Penn Treebank and tested on a set of sentences randomly generated from a probabilistic version of the grammar has on average 7.2×10^{27} parses per sentence according to Moore's work (Moore, 2000). Such a mammoth extent of result is also no exception in parsing of Czech (Smrž and Horák, 2000) (see Fig. 1) due to free word order and

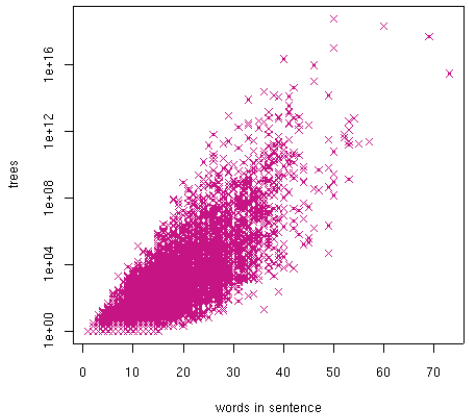

Figure 1: The dependence of number of resulting analysis on the number of words in the input sentence

rich morphology of word forms whose grammatical case cannot often be unambiguously determined.

A traditional solution for these problems is presented by probabilistic parsing techniques (Bunt and Nijholt, 2000) aiming at finding the most probable parse of a given input sentence. This methodology is usually based on the relative frequencies of occurrences of the possible relations in a representative corpus. "Best" trees are judged by a probabilistic figure of merit (FOM).

The term "figure of merit" is usually used to refer to a function that prunes implausible partial analyses during parsing. In this paper, we rather take figure of merit as a measure bounding the true probabilities of the complete parses.

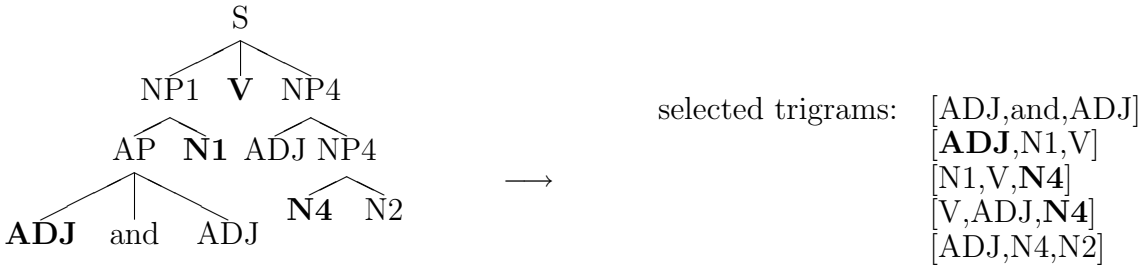

Figure 2: Lexical heads as n-gram's elements.

The standard methods of the best analysis selection (Caraballo and Charniak, 1998) usually use simple stochastic functions independent on the peculiarities of the underlying language. This approach seems to work satisfactorily in case of analytical languages. On the other hand, the obstacles brought by the synthetical languages in relationship with those simple statistical techniques are indispensable.

Therefore, we try to improve the standard FOMs taking into consideration specific features of free word order languages. The following text discusses the assets of three figures of merit that reflect selected phenomena of the Czech language.

2 Figures of Merit

The overall figure of merit of the syntactic analysis results is determined as a combination of several contributory FOMs that reflect particular language features such as

- frequency of syntactic constructs represented by pre-computed rule probabilities
- augmented n-gram model based on the occurrence of adjacent lexical heads standing for the corresponding subtrees
- affinity between constituents modeled by valency frames of verbs, adjectives and nouns

The selected FOMs participate on the determination of the most probable analysis. A straightforward approach lies in the linear combination of FOMs:

$$\xi = \lambda_1 \cdot \xi_1 + \lambda_2 \cdot \xi_2 + \lambda_3 \cdot \xi_3$$

where ξ_i are the FOMs' contributions and λ_i are empirically assigned weights (usually taken as normalizing coefficients). However, our experiments showed that the weights λ_i need to reflect the behaviour of particular lexical items, their categories or even analysed constituents. We thus need to handle the λ_i variables as functions of various parameters.

$$\xi = \lambda_1(\_) \cdot \xi_1 + \lambda_2(\_) \cdot \xi_2 + \lambda_3(\_) \cdot \xi_3$$

The following sections deal with the figures of merit that play a crucial role in the search for the best output analysis.

2.1 Rule-tied Actions and ξ_1 FOM

A key question is then what the good candidates for FOMs are. The use of probabilistic context-free grammars (PCFGs) involves simple CF rule probabilities to form a FOM (Chitrao and Grishman, 1990; Bobrow, 1991).

The evaluation of the first FOM is based on the mechanism of contextual actions built into the metagrammar conception (Smrž and Horák, 2000). It distinguishes four kinds of contextual actions, tests or constraints:

1. rule-tied actions
2. agreement fulfilment constraints
3. post-processing actions
4. actions based on derivation tree

The rule-based probability estimations are solved on the first level by the rule-tied actions, which also serve as rule parameterization modifiers.

Agreement fulfilment constraints are used in generating the expanded grammar (Smrž and Horák, 1999) or they serve also as chart pruning actions. In terms of (Maxwell III and Kaplan, 1991), the agreement fulfilment constraints represent the functional constraints, whose processing can be interleaved with that of phrasal constraints.

The post-processing actions are not triggered until the chart is already completed. The main part of FOM computation for a particular input sentence is driven by actions on this level. Some figures of merit (e.g. verb valency FOM, see Section 2.3) demand exponential resources for computation over the whole chart structure. This problem is solved by splitting the calculation process into the pruning part (run on the level of post-processing actions) and the reordering part, that is postponed until the actions based on derivation tree.

The actions that do not need to work with the whole chart structure are run after the best or n most probable derivation trees are selected. These actions are used, for example, for determination of possible verb valencies within the input sentence, which can produce a new ordering of the selected trees.

2.2 Augmented n-grams and ξ_2 FOM

The ξ_1 FOM is based on rule frequencies and is not capable of describing the contextual information in the input. A popular technique for capturing the relations between sentence constituents is the n-gram method, which takes advantage of a fast and efficient evaluation algorithm.

For instance, (Caraballo and Charniak, 1998) presents and evaluate different figures of merit in the context of best-first chart parsing. They recommend boundary trigram estimate that has achieved the best performance on two testing grammars. This technique, as well as stochastic POS tagging based on n-gram statistics, achieves satisfactory results for analytical languages (like English). However, in case of free word order languages, current studies suggest that these simple stochastic techniques considerably suffer from the data sparseness problem and require a huge amount of training data.

The reduction of the number of possible training schemata, which correctly keeps the correspondence with the syntactic tree structure, is achieved by elaborate selection of n-gram candidates. While the standard n-gram techniques work on the surface level, this approach allows us to move up to the syntactic tree level. We advantageously use the ability of *lexical heads* to represent the key features of the subtree formed by its dependants (see Figure 2). The principle of lexical heads has shown to be fruitfully exploited in the analysis of free word order languages. The obtained cut-down of the amount of training data may be also crucial to the usability of this stochastic technique.

2.3 Verb Valencies and ξ_3 FOM

Our experiments have shown that, in case of a really free word order language, the FOMs ξ_1 and ξ_2 are not always able to discover the correct reordering of analyses. So as to cope with the above mentioned difficulties in Slavonic languages (namely Czech), we propose to exploit the language specific features. Preliminary results indicate that the most advantageous approach is the one based upon valencies of the verb phrase — a crucial concept in traditional linguistics.

The part of the system dedicated to exploitation of information obtained from a list of verb valencies (Pala and Ševeček, 1997) is necessary for solving the prepositional attachment problem in particular. During the analysis of noun groups and prepositional noun groups in the role of verb valencies in a given input sentence one needs to be able to distinguish free adjuncts or modifiers from obligatory valencies. We are testing a set of heuristic rules that determine

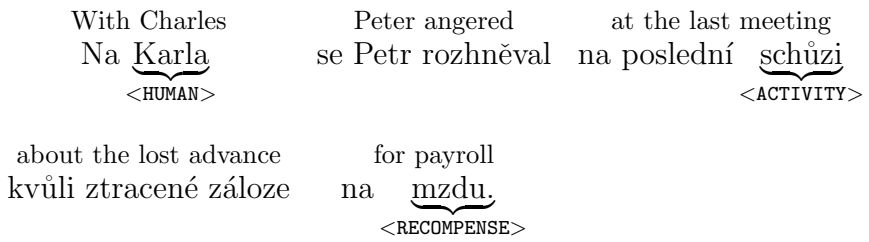

Figure 3: Free adjuncts identification by means of lexico-semantic constraints.

whether a found noun group typically serves as a free adjunct. The heuristics are based on the lexico-semantic constraints (Smrž and Horák, 1999).

An example of the application of the heuristics is depicted in Figure 3. In the presented Czech sentence, the expression `na Karla` (with Charles) is denoted as a verb argument by the valency list of the verb `rozhněvat se` (anger), while the prepositional noun phrase `na schůzi` (at the meeting) is classified as a free adjunct by the rule specifying that the preposition `na` (at) in combination with an `<ACTIVITY>` class member (in locative) forms a location expression. The remaining constituent `na mzdu` (for payroll) is finally recommended as a modifier of the preceding noun phrase `záloze` ([about the] advance).

Certainly, we also need to discharge the dependence on the surface order. Therefore, before the system confronts the actual verb valencies from the input sentence with the list of valency frames found in the lexicon, all the valency expressions are reordered. By using the standard ordering of participants, the valency frames can be handled as pure sets independent on the current position of verb arguments.

2.4 Preferred Word Order

In analytical languages, the word order is usually taken as rather fixed and that is why it can be employed in parsing tree pruning algorithms. However, in case of inflectional languages, the approaches to word order analysis are diverse. The most influential theory works with the topic-focus articulation (Sgall et al., 1986). Although nearly all rules that could limit the order of constituents in Czech sentences can be fully relaxed, a standard order of participants can be defined. A corpus analysis of general texts affirms that this *preferred word order* is often followed and that it can be advantageously used as an arbiter for best analysis selection.

Cases where the ξ_i FOMs do not unambiguously elect the best candidates can be routed by the preferred word order in the form of functional weights $\lambda_i(\_)$ with appropriate parameters.

3 Results

This section presents results of experiments with the stated figures of merit for the best analysis selection algorithm. First, the acquisition of training data set derived by exploitation of a standard dependency tree bank for Czech is described. Then, we step to a comparison of parser running times with that of another available parser.

3.1 The Training Set Acquisition

A common approach to acquiring the statistical data for analysis of syntax employs learning the values from a fully tagged tree bank training corpus. Building of such corpora is a tedious and expensive work and it requires a team cooperation of linguists and computer scientists. At present the only source of Czech tree bank data is the Prague Dependency Tree Bank (PDTB) (Hajič, 1998), which includes dependency analyses of about 100 000 Czech sentences.

First, in order to be able to exploit the data from PDTB, we have supplemented our grammar with the dependency specification

precision on sentences	percentage
of 1-10 words	86.9 %
of 11-20 words	78.2 %
of more than 20 words	63.1 %
overall precision	**79.3 %**
number of sentences with mistakes in input	8.0 %

Table 1: Precision estimate

for constituents. Thus the output of the analysis can be presented in the form of pure dependency tree. In the same time we unify classes of derivation trees that correspond to one dependency structure. We then define a canonical form of the derivation to select one representative of the class that is used for assigning the edge probabilities.

This technique enables us to relate the output of our parser to the PDTB data. However, the profit of exploitation of the information from the dependency structures can be higher than that and can run in an automatically controlled environment. For this purpose, we use the mechanism of *pruning constraints*. A set of strict limitations is given to the syntactic analyser, which passes on just the compliant parses. The constraints can be either supplied manually for particular sentence by linguists, or obtained from the transformed dependency tree in PDTB.

The Table 1 summarizes the precision estimates counted on real corpus data. These measurements presented here may discount the actual benefits of our approach due to the estimated 8% of mistakes in the input corpus.

3.2 Running Time Comparison

The effectivity comparison of different parsers and parsing techniques brings a strong impulse to improving the actual implementations. Since there is no other generally applicable and available NL parser for Czech, we have compared the running times of our syntactic analyser on the data provided at `http://www.cogs.susx.ac.uk/lab/nlp/carroll/cfg-resources/`.

These WWW pages resulted from discussions at the Efficiency in Large Scale Parsing Systems Workshop at COLING'2000, where one of the main conclusions was the need for a bank of data for standardization of parser benchmarking. The best results reported on standard data sets (ATIS and PT grammars) until today are the comparison data by Robert C. Moore (Moore, 2000). In the package, only the testing grammars with input sentences are at the disposal, the release of referential implementation of the parser is currently being prepared (Moore, personal communication).

ATIS grammar, Moore's LC_3 + UTF	11.6
ATIS grammar, our system	7.2
PT grammar, Moore's LC_3 + UTF	41.8
PT grammar, our system	57.2

Table 2: Running times comparison (in seconds)

Since we could not run the referential implementation of Moore's parser on the same machine, the above mentioned times are not fully comparable (we assume that our tests were run on a slightly faster machine than that of Moore's tests). We prepare a detailed comparison, which will try to explain the differences of results when parsing with grammars of varying ambiguity level.

4 Conclusions

The methods of the best analysis selection algorithm described in this paper show that the parsing of inflectional languages calls for sensitive approaches to the evaluation of the appropriate figures of merit. The case study of Czech suggests that the use of language specific features can improve the results of simple stochastic techniques on annotated corpus data.

Future directions of our research lead to improvements of the quality of training data set so that it would cover all the most frequent language phenomena. Our investigations indicate that, in addition to verbs, the best analysis selection algorithms could also

take advantage of valency frames of other POS categories (nouns, adjectives).

References

R. J. Bobrow. 1991. Statistical agenda parsing. In *Proceedings of the February 1991 DARPA Speech and Natural Language Workshop*, pages 222–224. San Mateo: Morgan Kaufmann.

H. Bunt and A. Nijholt, editors. 2000. *Advances in Probabilistic and Other Parsing Technologies*. Kluwer Academic Publishers.

S. Caraballo and E. Charniak. 1998. New figures of merit for best-first probabilistic chart parsing. *Computational Linguistics*, 24(2):275–298.

M. Chitrao and R. Grishman. 1990. Statistical parsing of messages. In *Proceedings of the Speech and Natural Language Workshop*, pages 263–266, Hidden Valley, PA.

J. Hajič. 1998. Building a syntactically annotated corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning*, pages 106–132, Prague. Karolinum.

J. T. Maxwell III and R. M. Kaplan. 1991. The interface between phrasal and functional constraints. In M. Rosner, C. J. Rupp, and R. Johnson, editors, *Proceedings of the Workshop on Constraint Propagation, Linguistic Description, and Computation*, pages 105–120. Instituto Dalle Molle IDSIA, Lugano. Also in Computational Linguistics, Vol. 19, No. 4, 571–590, 1994.

R. C. Moore. 2000. Improved left-corner chart parsing for large context-free grammars. In *Proceedings of the 6th IWPT*, pages 171–182, Trento, Italy.

K. Pala and P. Ševeček. 1997. Valencies of Czech verbs. In *Proceedings of Works of Philosophical Faculty at the University of Brno*, pages 41–54. Brno. (in Czech).

P. Sgall, E. Hajičová, and J. Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

P. Smrž and A. Horák. 1999. Implementation of efficient and portable parser for Czech. In *Text, Speech and Dialogue: Proceedings of the Second International Workshop TSD'1999*, Pilsen, Czech Republic. Springer Verlag, Lecture Notes in Computer Science, Volume 1692.

Pavel Smrž and Aleš Horák. 2000. Large scale parsing of Czech. In *Proceedings of Efficiency in Large-Scale Parsing Systems Workshop, COLING'2000*, pages 43–50, Saarbrucken: Universitaet des Saarlandes.