The Linguistic Basis of a Rule-Based Tagger of Czech[*]

Karel Oliva[1], Milena Hnátková[2], Vladimír Petkevič[2], and Pavel Květoň[2]

[1] Computational Linguistics, University of Saarland, Saarbrücken, Germany
oliva@coli.uni-sb.de
[2] Faculty of Arts, Charles University, Prague, Czech Republic
{Milena.Hnatkova,Pavel.Kveton,Vladimir.Petkevic}@ff.cuni.cz

Abstract. This paper describes the conception of a rule-based tagger (part-of-speech disambiguator) of Czech currently developed for tagging the *Czech National Corpus* (cf. [2]). The input of the tagger consists of sentences whose words are assigned all possible morphological analyses. The tagger disambiguates this input by successive elimination of tags which are syntactically implausible in the sentential context of the particular word. Due to this, the tagger promises substantially higher accuracy than current stochastic taggers for Czech. This is documented by the results concerning the disambiguation of the most frequent ambiguous word form in Czech – the word *se*.

1 Introduction

An automatic morphological disambiguation of large textual corpora is one of the main tasks in contemporary corpus linguistics. The majority of corpora are in English and they are morphologically disambiguated by stochastic taggers with accuracy over 97 %. Unlike English, Czech is characterized by a high degree of free-word order and by the absence of the abundant usage of unambiguous "small" words (articles, auxiliary verbs, etc.). Hence in Czech a stochastic tagger, selecting the most likely tag for a word by considering the immediately neighbouring words, makes often incorrect choice because it cannot base its decision on the dense sentence skeleton induced by the fixed word order in combination with frequent occurrences of functional words (as is the case in English).

On the contrary, the decisions of a rule-based tagger (in the spirit of [8]) can be based on the entire sentential context and hence they can very often profit from the rich inflection of Czech.

2 Assessment of the Stochastic Disambiguation of Czech

The highest correctness rate hitherto achieved by statistical-based tagging for Czech is 93,85 % (cf. [6]). This number is to be understood, however, as resulting

[*] The work described is funded by the *GAČR* grant No. 405/96/K214.

from the combination of performance of the stochastic tagger used *and* the ambiguity of the text. As about 66 % of the tokens in the *Czech National Corpus* are unambiguous (i.e. the tagger has nothing to decide on these tokens), the actual success rate of the tagger is in fact about 71 %. After careful inspection of a portion of stochastically disambiguated data we can summarize the main reasons for such a state of affairs as follows:

- stochastic methods seem to be very primitive when applied on a syntactically complex, free word-order language without an abundance in syntactically fixed points (such as Czech);
- stochastic taggers are dependent on the training data. If they have not "seen" a particular configuration of tags in the training corpus, they cannot cope with this configuration properly. In other words, stochastic taggers model *parole* (in the de Saussurean wording) rather than *langue*. As there are immense quantities of distinct complex manifestations of phenomena (esp. due to their different combinations and word order variants) in any Czech text, the modelling of *parole* – unlike the modelling of *langue* – cannot be adequate. In other words, the source of many problems of stochastic tagging is its underlying assumption that texts of a language can be viewed as accidental sequences of word forms, an assumption standing in opposition to the obvious fact that natural languages are systems *sui generis*;
- stochastic taggers are dependent on the size and the repertory of tags. This dependency turns to be fatal: the more fine-grained input the morphological analysis provides (i.e. the more refined information is contained in the tagset), the worse the overall results are. Closely related to this is also the fact that a stochastic tagger for a flective language has to use a large tagset (1100+ tags are reported in [5]) and hence it should be trained on a very large manually disambiguated corpus (millions of word forms), but creating such a corpus is infeasible in practice;
- it is very difficult to localize and fix the sources of errors which a stochastic tagger commits (to *debug* the tagger); these taggers are black boxes driven only by data they encountered, and as such they often behave contrary to expectations. For one instance out of many: in Czech a preposition can never be followed by a word form in the nominative case but the most successful stochastic tagger of Czech sometimes does disambiguate such a word form as nominative (though the sequence could not occur in the training corpus!);
- stochastic taggers of Czech do not make use of information about the structure of language which is, however, available in the data. For instance, they do not use the systemic phenomenon consisting in the vocalization of those occurrences of prepositions that are immediately followed by word forms beginning with specific consonant clusters.

As a matter of fact, we consider all these points to be true in general, not only in the particular case of Czech – even though they get a substantial confirmation on Czech as a test case. The conclusions which we draw from this are that the morphological disambiguation of Czech must be performed in such a way that it will have none of the above-mentioned drawbacks of currently available stochastic taggers of Czech, and, fundamentally, that the disambiguation must be based on exploiting as much linguistic information provided by the system of language as possible.

3 General Conception of the Rule-Based Tagger

The core idea of the rule-based tagging process is the combination of a morphological analyzer (lemmatizer) with linguistically-based tag-elimination rules, which has been presented for the first time in [8]. In particular, the point is that for a language with rich inflection such as Czech, the morphological analysis of the text is an indispensable basis of any tagging (even a statistics-based one, cf. [5]). To put things clearly, we consider the morphological analysis to be a process assigning each word form in the text the set of all possible tags appurtenant to this word form, irrespectively of its environment (neighbouring words) as well as of the probability of an occurrence of the particular form with the particular tag (i.e. word form with its "morphological meaning") in a real text (as an example of an intuitively very unrealistic – i.e. improbable – meaning, let us take the word form *čas* as the form of the imperative of the verb *časiti (se)*). For this purpose, we use the morphological analyzer and a tagset developed by Jan Hajič (cf. [3], [4]).

In the entire tagging process, the morphological analysis (i.e. multiple tag assignment) is then followed by tag elimination, that is, by a set of procedures (rules) trying to narrow down the set of tags associated (by the morphological analysis) with a particular word form, ideally achieving the situation in which only the correct tag is associated with each word of the input. Unlike morphological analysis, this elimination makes use of context of the word form whose tags are to be eliminated, and it is, in fact, based on this context.

Provided the morphological analysis has been already implemented, the true task of rule-based tagging consists in the development of the tag-elimination procedures.

As the first step in solving this task, we studied the types of morphological ambiguity occurring in Czech. Classified broadly, there are two types of such ambiguity: provisionally, we call them *regular (paradigm-internal)* and *casual (lexical)*.

The regular (paradigm-internal) ambiguities are those which occur within a paradigm, i.e. which are common to all lexemes belonging to a particular inflection class. As an example, we might take the paradigm of masculine inanimate nouns with stems ending in a non-soft consonant (declension pattern *hrad*), with systemic ambiguity among nominative, accusative and instrumental case of plural number (e.g., *hrady*). (In this respect the inflection is to be taken rather broadly, including, e.g., regular derivation of verbal nouns and adjectives, cf. the example of the word form *stavění*, where the ambiguity goes also between a nominal and an adjectival reading.)

Massive as this kind of ambiguity is, it is relatively easy to discover and classify since it is paradigmatical and hence it can be read off the inflection tables of Czech morphology.

More intricate is the study of the casual (lexical, paradigm-external) morphological ambiguity. This ambiguity is lexically specific and hence cannot be investigated via paradigmatics. Nevertheless, a detailed knowledge of this ambiguity is essential for the aim of tagging, since only after grasping the ambiguity classes, it becomes clear *which* ambiguities are to be resolved, and strategies can be developed *how* to do this. As a result of work in this field we arrived at a listing of approximately 125 classes of such ambiguous word forms (where a class is constituted by word forms displaying the same kind of ambiguity, e.g., the forms which have a reading as a noun in a particular case and number, and a verbal reading, or nominal feminine and masculine reading, etc.), apart from hundreds of forms which are, as a rule, unique (the "class" has just one member) but display triple or more ambiguity[1].

The general idea following from this is then that the rules of the tagger are organized into packages, each aimed at resolving one ambiguity class by stepwise discarding the contextually impossible tags of a word (cf. again [8]). The set of all (packages of) elimination rules we use can be classified according to two criteria: according to locality/nonlocality of their operation and according to the level of reliability a rule can be assigned.

The most important feature of the approach is the possibility for the tag elimination process to operate on a context whose range is not fixed in advance. The beneficial impact of this becomes clear in comparison with both the statistical taggers and the Brill taggers ([1]), where in both cases the limitation of the operational scope of the tagger to a local context (window of a fixed size) seems to be the source of most errors.

The importance of non-locality, obvious even for English, becomes truly crucial for languages with so-called "free word-order" (such as Czech). Thus, for example, in deciding the status of the word *se* which is ambiguous between a preposition and a reflexive particle/pronoun, a rule-based tagger can take into consideration the presence/absence of a reflexive tantum verb in the sentence (in any position of the sentence, not just in the immediate neigbourhood of *se*), which brings along a decisive advantage over the locally operating methods which commit errors in this point exactly due to the lack of globality.

However, obviously nothing prevents us also from the use of purely local linguistic strategies: to take another example from the package of rules devoted to solving the ambiguity of *se*, it is a linguistically trivial fact that the prepositional reading of *se* can be available only in local contexts requiring the vocalization of the basic form of the preposition *s* resulting in the form *se*.

Another profound conceptual difference between a rule-based tagger and a stochastic one is that the rules (and hence also results) of the former can be

[1] The highest level of lexical ambiguity, fivefold, which has been discovered is represented by the word *kose*; together with paradigmatical ambiguity, this adds up to seven different morphological analyses of this word form.

assigned reliability (either as binary opposition "fully reliable"/"not fully reliable" or on a more fine-grained scale, discrete or continuous[2]) which seems to be in principle impossible with the latter (this is to say, the results of a stochastic tagger can never be claimed as fully reliable).

It is obvious that it is desirable to perform the tagging with as many fully reliable rules as possible. As a matter of fact, our rules are currently divided into two classes ("fully reliable"/"not fully reliable") as follows:

- all tag-elimination rules reflecting general syntactic regularities of Czech belong to the class of fully reliable rules (not only wrt. some testing corpus, but generally)
- apart from these general rules, we have implemented a number of rules coping with idioms and collocations; since it is (next to) always possible to construct an example where the construction otherwise interpreted as an idiom or a collocation can also get a literal meaning, this module cannot be considered as fully reliable[3].

4 First Implemented Results

Although the work on the tagger started only in spring this year, we can already present the first results and compare them to those obtained by the stochastic and Brill-like taggers developed for Czech ([5], [6]).

The first package of rules created concerns the disambiguation of the word form *se* (ambiguous between a reflexive particle/pronoun and a preposition), since this is the most frequent ambiguous form in Czech (according to [7], *se* is the eighth most frequent Czech word, and the most frequent one among all ambiguous words; in the corpus formed by the current Czech translation of George Orwell's *1984* it is the most frequent Czech word form at all – cf. [9]). After implementing slightly more than twenty general syntactic rules concerning *se*, we arrived at disambiguating 94.34 % of occurrences of *se* (700 out of 742) within a sample text of 39.000+ words (one issue of the scientific magazine *Vesmír* dated 1992). These figures are to be interpreted as follows: 94.34 % of occurrences of *se* are decided correctly, while the rest, i.e. 5.66 %, remains ambiguous[4]. In the particular task of the disambiguation of *se*, the incorporation

[2] Currently we use only the binary scale, and we set the reliability manually – however, there is work in progress which should result in a continuous scale of reliability of rules measured automatically by the percentage of errors a rule makes when resolving ambiguity within a testing corpus.

[3] Let us remark, however, that in the body of the *Czech National Corpus* comprising 100.000.000+ words we did not encounter a construction where an incorrect tag would be assigned by using this module (and hence it is obvious that the counterexamples are of only rather theoretical nature).

[4] This is to be compared with the statistical tagger currently in experimental use for tagging the *Czech National Corpus*, where the success rate of resolving the ambiguity of *se* ranks at the level of about 61 % of occurrences, while the remaining cca 39 % are not left unresolved but are disambiguated incorrectly.

of the module processing idioms and collocations did not bring much progress in this case, as the number of disambiguated occurrences of *se* rose to 702 only, thus yielding the total of 94.60 % success rate and leaving 5.40 % unresolved. However, after inspecting the undecided cases we are developing additional rules, with the prospect of arriving finally at the stage where only genuinely syntactically ambiguous cases are left undecided at last.

5 Remarks on Computer Implementation

The system (i.e. the rules and the software environment needed for their application) is currently implemented in C++ under the Linux operating system. In particular, the rules are created by linguists in a semi-formal format and only then programmed in C++. This makes the implementation of a rule slow and above all error-prone as well as slightly difficult to debug (redecoding C++ program code back to a "human" language is not entirely simple). Therefore we are currently developing a "rule language", i.e. a language in which the linguist can write the rules directly. The development of such a rule language involves the definition of the language itself and the interpreter or compiler of the rules transforming them to their executable form.

References

1. Brill, E.: A Simple Rule-Based Part-of-Speech Tagger. Proceedings of the Third Conference on Applied Natural Language Processing. Trento (1992)
2. Czech National Corpus. Faculty of Arts, Charles University. http://ucnk.ff.cuni.cz
3. Hajič, J.: Unification Based Morphology Grammar. PhD Thesis. MFF UK (1994)
4. Hajič, J.: Morfologické značky pro užití v Českém národním korpusu. ms.
5. Hajič, J., Hladká, B.: Probabilistic and Rule-Based Tagger of an Inflective Language – a Comparison. Proceedings of the Fifth Conference on Applied Natural Language Processing. Washington D.C. (1997)
6. Hladká, B.: Czech Language Tagging. PhD Thesis. MFF UK (2000)
7. Jelínek, J., Bečka, J. V., Těšitelová, M.: Frekvence slov, slovních druhů a tvarů v českém jazyce. Praha (1961)
8. Karlsson, F., Voutilainen, A., Heikkilä, J., Antilla, A. (eds.): Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin New York (1995)
9. Petkevič, V.: Korpus románu George Orwella '1984'. In prep.