Rule-Based Tagging: Morphological Tagset versus Tagset of Analytical Functions

Kiril Ribarov

Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Prague
Czech Republic

e-mail: ribarov@ufal.mff.cuni.cz

Abstract

This work presents a part of a more global study on the problem of parsing of Czech and on the knowledge extraction capabilities of the Rule-based method. It is shown that the successfulness of the Rule-based method for English and its unsuccessfulness for Czech, is not only due to the small cardinality of the English tagset (as it is usually claimed) but mainly depends on its structure ("regularity" of the language information).

1. Background

Within the Prague Linguistic School, the language is understood as a system of layers, where each layer by itself is a system with many relations and its own semantics, while the layers are interrelated by homomorphic mappings. Opposed to generation, the analysis is a process of giving a form its meaning, always between two adjacent layers in the order: phonetic, morphonological, morphematic, syntactic, tectogrammatical (Sgall, Hajičová, Panevová, 1986). Thus, in order to move from a 'lower' layer to an 'upper' one, we rely on the knowledge of the lower layers, from which an additional knowledge has to be extracted.

The Prague Dependency Treebank (PDT), in relation to the aforementioned theory, has a three-level structure (Hajič 1998):

- full morphological tagging, the lowest level,
- syntactic annotation using dependency syntax, where each token is given a so-called analytical function,
- tectogrammatical level, the highest level.

In the sequel the second level structure will be considered.

2. Analytical Fuctions

Besides the dependency tree structure assigned to each sentence, the second-level of annotation of the PDT, assignes a surfice syntactic attribute to each node of the dependency tree called an analytical function (afun). Each token (including punctuation marks) represents a separate node. No extra nodes are being added neither tokens are being deleted, except for an extra node being added in order to mark the whole sentence (the tree root).[1].

[1]The second level of annotation is also called an analytical level.

The analytical functions include tags for predicate, subject, object, adverbial, various complements, attribute, auxiliaries, reflexives, conjunctions, prepositions; relation of coordination, apposition and parenthetical expressions. Ellipsis are also handeled. A complete list of the main analytical functions are presented in Table 1.

Each afun (except for AuxX, AuxG, AuxS and AuxK) may have one of the following three suffixes: *_Co (for coordination), *_Ap (for apposition) or *_Pa (for paranthetical expression).

3. Specification

While in the case of morphological tagging, it is clear that each token is directly assigned its morphological description, on the analytical level the process of annotation is not that straightforward. Several approaches were tested, which can be classified into methods which:

(a) firstly capture the dependency structure (Hajič, Ribarov 1997) and (Ribarov, 1996), then, the analytical functions are added,

(b) firstly assign the analytical functions and then assign the dependency structure,

(c) assign both, the analytical functions and the dependency structure, simultaneously.

Although the variant (b) may seem less reasonable than the others, all of them end with almost the same success rate (at the moment). The variant (b) is interesting from the aspect that its first stage, the assignment of analytical functions, can be transformed to a classical problem of annotation, using the methods heavily used for morphological tagging. Further, if the analytical functions are given, succesful reconstruction of the dependency tree is highly probable.

Therefore, in the following, we will devote ourselves to the problem of annotation related to the structure of the annotation (tag) sets from the point of view of a success rate,

AFUN	EXPLANATION
Pred	Predicate if it depends on the tree root
Sb	Subject
Obj	Object
Adv	Adverbial
Atv	Complement
AtvV	Complement, if only one governor is present
Atr	Attribute
Pnom	Nominal predicate's nominal part, depends on the copula "to be"
AuxV	Auxiliary Verb "to be"
Coord	Coordination node
Apos	Apposition node
AuxT	Reflexive particle, lexically bound to its verb
AuxR	Reflexive particle, which is neither Obj nor AuxT (passive)
AuxP	Preposition or a part of compound preposition
AuxC	Subordinate conjunction
AuxO	(Superfluously) referring particle or emotional particle
AuxZ	Rhematizer or another node acting to another constituent
AuxX	Comma, but not the main coordinating comma
AuxG	Other graphical symbols bing not classified as AuxK
AuxY	Other words, such as particles without a specific syntactic function, parts of lexical idioms, etc.
AuxS	The added root of the tree
AuxK	Punctuation at the end of the sentence or direct speech or citation clause
ExD	Ellipsis handling: function for nodes which "pseudo depend" on a node on which they would not depend if there were no ellipsis
AtrAtr, AtrAdv, AdvAtr, AtrObj, ObjAtr	Two-part afun; a node which could depend also on its governor's governor and thus have the appropriate second fuction. There must be no semantical or situational difference between the two cases. The second function represents the annotator's preference.

Table 1: Analytical Functions in the PDT

under the scope of the Rule-based approach (Brill, 1993).

Previous work of a similar kind but restricted only to the morphological tagset (Hladká, Ribarov 1998), shows that the success rate of a POS tagging algorithm depends on the cardinality of the tagset. Experiments with variously sized morphological tagsets of Czech, when the Rule-based method is applied gives the following success rates[2]:

- for a tagset of 1171 morphological tags a success rate of approx. 80%, and
- for a tagset of 206 morphological tags a success rate of approx. 87%.

[2]For the success rates are still not satisfactory, the success rates are given in approximate values.

We would like to note that the morphological tagging of Czech (as a highly flective language), has a high level of ambiguity. Further, there are several morphological tagsets for Czech (more or less detailed) including the full tagset of more than 3000 tags[3].

4. Rule-Based afun Tagging

Proper usage of the Rule-based tagger claims no dependence on a tagset, which allows one, to change the tagset, so instead of tagging the tokens by tags from the morphological tagset, to use tags from the set of analytical functions.

In terms of its cardinality the set of analytical functions is rather small resulting in 70 tags.

If applied as such and tested, without using any additional information, the Rule-based tagger (trained on the pair: lexical token, afun) yields a success rate of approx. 66% on the test set for the analytical functions.

Since the syntactic level is superordinate to the morphematic one, and in order to trace the reason for such a low success rate (when compared to the success rate when the Rule-based tagger is used in order to tag the tokens by their morphological tags), in the next experiment, information of the morphematic level, i.e. the morphological tags, was included. At this step, the lexical token has been substituted by its morhological tag. Thus, the Rule-based tagger was trained on the pair (disambiguated morphological tag, analytical function).

In this case a success rate of 72% was recorded, which is only 6% improvement compared to the 66% when the morphological information was not taken into consideration.

Both of the experiments were provided on the same training sets, and tested on the same test sets, hence the results are directly comparable.

The aim of the presented success rates is not to claim the highest possible success rates for the given applications, but the success rates of the Rule-based tagger when being directly applied without any (outside of the core algorithm) improvement.

5. Conclusion

Without performing the experiments it would be difficult to expect such results, since it would be expected that the usage of the morphological information, would contribute more significantly to the success rate of tagging of the analytical functions.

One may conclude that:

- The success rate of the Rule-based tagger depends not only on the cardinality of the tag set but even more on its structure. Hence, the claim that a higher success rate is due to low cardinality of the tagset, when success rates of the same task but over different languages are being compared, is not sufficient.

[3]On morphological tagging of Czech see (Hladká, 2000).

- The Rule-based tagger, as originally proposed by Brill (Brill, 1992), cannot be directly applied for a successful tagging of analytical functions.

- One of the basic reasons, as shown by analysis of the PDT, is in the different nature of homonymy present within analytical functions. Other reasons include, as Czech is free order language, that it is frequently not the case that the immediate neigbouring tokens determine the value of the token to be tagged.

6. Acknowledgement

The results described herein have been obtained within the Ministry of Education Project No. VS96151.

7. References

E. Brill. A Simple Rule-Based Part of Speech Tagger. In: *Proceedings of the Third Conference of Applied Natural Language Processing*. Trento, Italy. 1992.

E. Brill. *A Corpus-Based Approach to Language Learning*. A dissertation in Department of Computer and Information Science, University of Pennsylvania, Philadelphia, USA, 1993.

J. Hajič. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. *Issues of Valency and Meaning (Studies in honour of Jarmila Panevova)*. Karolinum, Charles University Press, Prague, Czech Republic, 1998.

J. Hajič and K. Ribarov. Rule-Based Dependences. In: *Proceedings of the Workshop on Empirical Learning of Natural Language Processing Tasks, MLNet, Prague, Czech Republic, April 23-25, pp.125-136, 1997.*

B. Hladká. *Czech Language Tagging*. PhD Thesis at the Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, 2000.

B. Hladká and K. Ribarov. Part of Speech Tags for Automatic Tagging and Syntactic Structures. *Issues of Valency and Meaning (Studies in honour of Jarmila Panevova)*. Karolinum, Charles University Press, Prague, Czech Republic, 1998.

K. Ribarov. *Automatická tvorba gramatiky přirozeného jazyka*. MSc dissertation in Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 1996.

P. Sgall, E. Hajičová and J. Panevová. *The Meaning of the Sentence and its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, Holland, 1986.