Tools for Semi-Automatic Assignment of Czech Nouns to Declination Patterns

Dita Bartůšková and Radek Sedláček

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
ydita@aurora.fi.muni.cz,rsedlac@fi.muni.cz

Abstract. In this paper, we present tools for semi-automatic assignment of Czech nouns to declination patterns. First, we explain the reasons for development of such tools and then we describe the structure of the system in detail. It is based on a decision tree that consists of questions and answers allowing to distinguish particular declination patterns. Finally, we provide basic statistic data that clarify the relation between the patterns we developed and the classical ones.

1 Introduction

One of the most time and money consuming phases in the process of the software development is its maintenance [1]. This holds true also in the case of programs occuring in the field of natural language processing. Moreover, in the systems that are based on various dictionaries or databases, there is one more complication, i.e. the maintenance of the stored data. Typical example of such a specialised database is a dictionary used by a morphological analyser (e.g. [2–5]). Dynamic actualisation (and the extension in particular) of the dictionary is important for increasing the number of successfully recognised word forms by the analyser in real texts. Furthermore, the correctness of this action (the assignment of new words to corresponding declination patterns) is crucial for further error-free processing. Due to this, the reason for automatising this process as much as possible is well-founded.

2 Structure of the System

The system we developed is based on interaction with an experienced user (usually an expert or a linguist) and the whole process of adding new words to particular declination patterns is a written answer-question dialogue. The hierarchical system of patterns based on distinctive features allows such an order of questions that leads to the precisely determined pattern. The aim of the system is to minimise the number of interactions, i.e. to minimise the number of questions that cannot be answered automatically by the system (e.g. on the base of the formal form of the new word), but they have to be answered by an expert.

We have implemented two program tools: a tool for user-friendly editing of the decision trees and an interpreter which can parse the trees and interacts with the user via answer-question dialogue. Indexing techniques and data structures [6, 7] used in the implementation allow an efficient storage and retrieval of the data.

2.1 Decision Tree

The basic part of the system is the pattern hierarchy, strictly speaking the structured set of questions reflecting the distinctive features of the particular patterns. We have elaborated the hierarchical system of declination patterns for Czech nouns, nevertheless the data structures and the formalism used for encoding questions and answers are applicable for other parts of speech as well.

To be precise, the hierarchy of questions is represented by a special form of an n-ary decision tree [8]. In the internal nodes, there are `answer-question` pairs where `answer` corresponds to the `question` stored in the predecessor. It is clear that in the root of the tree there is no answer and in the leaves there is no question. In the leaves, there are names of appropriate declination patterns instead of questions. We use the "?" character to distinguish questions from pattern names in the leaves. The following figure 1 shows an example of a decision tree.

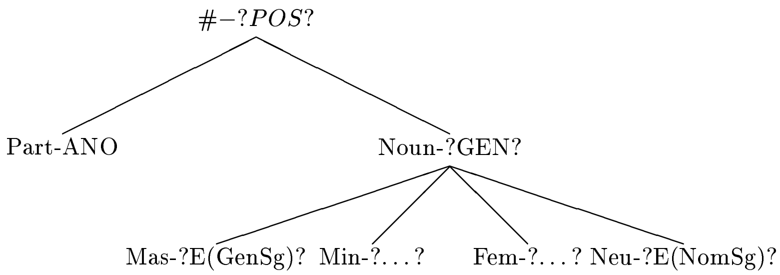

Fig. 1. Decision tree

2.2 Interpreter

The interpreter parses a decision tree from its root. It asks the first question and waits for the answer (if it is not able to answer it itself). Once it gets the answer, it goes on to the successor in the tree dependent on the answered value. Then it asks the next question or determines the name of the pattern (if it is in the leaf).

Questions and answers are for the purpose of automatic processing encoded into a compact form, but it is still the possible to translate them into the user's natural language (e.g. "?E(GenSg)?" can stand for "*What is the ending in genitive singular?*" in the above figure 1). Answers can be then viewed as values from a pre-defined domain. In fact, every question has its domain of possible answers and the user chooses the right one.

3 Declination Patterns

While creating the decision tree for declination patterns of Czech nouns, we bear in mind the following general principles:

1. Exceptions (or language phenomena) are dealt with as soon as possible, i.e. we try to eliminate them just at the beginning of the decision process. In most cases, exceptions are indicated automatically (e.g. foreign names, words with Latin endings etc.).
2. We try to keep the relation between the patterns we developed and the classical ones [9], however, there are some patterns in the system that fluctuate between two classical patterns.
3. Patterns that contain a certain type of vocalic alternations dependent on the type of the sounds of speech (e.g. consonant/vocal, hard/soft), are situated on the same level.
4. Answers that have to be answered by the user are dealt with in the later phases of the decision process; we use automatically answerable questions first.

For example, figure 2 shows which word forms (in which cases) are distinctive of masculines inanimate with the appropriate ending in genitive singular.

```
-u   --> NomSg --> LocPl --> LocSg
               --> LocSg --> LocPl
-u|a --> NomSg --> LocSg --> LocPl
               --> LocPl
-a   --> NomSg
     --> LocSg --> LocPl
-e   --> NomSg --> LocPl
     --> LocSg --> LocPl
     --> NomSg
-ě   --> NomSg
-e|u --> LocSg --> NomSg
-a|e --> LocSg
```

Fig. 2. Distinctive features of masculines inanimate

We can see that the four questions are enough to determine the corresponding pattern. The relation to the classical patterns (cf. [10]), the number of exceptions and the maximum number of levels in the respective decision tree for all Czech nouns is provided in table 1.

Table 1. Relation to classical patterns

Masc. an.	**#patt.**	**Masc. in.**	**#patt.**	**Femin.**	**#patt.**	**Neut.**	**#patt.**
pán	44+28	hrad	49+14	žena	82+17	město	41+19
muž	22+2	les	15+1	růže	18+3	moře	5
předseda	15+3	stroj	17	píseň	9	kuře	7
soudce	2	hrad/les	14+1	kost	15	stavení	2
pán/muž	4	les/stroj	2	žena/růže	6+1	město/moře	3+1
		stroj/hrad	6	píseň/kost	18		
exceptions	6		3		2		8+3
ind./adj./pl.t.	2+8+5		2+1+28		1+4+29		7+4+11
total	141		153		205		111
levels	5		4		5		6

4 Conclusion

We have presented a tool that is used for semi-automatic assignment of new words to their respective declination patterns. The core of the presented system consists in a decision tree which encodes the distinctive features of particular patterns. At present, we have worked out a tree for Czech nouns, but our aim is to apply the same principle to other inflectional parts of speech, particularly to adjectives and verbs.

References

1. Sommerville, I.: Software engineering. 5th edn. Addison Wesley, Wokingham (1996)
2. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). 1st edn. Karolinum Press, Praha (2001)
3. Lezius, W., Rapp, R., Wettler, M.: A Freely Available Morphological Analyzer, Disambiguator and Context Sensitive Lemmatizer for German. In: Proceedings of the COLING-ACL. (1998)
4. Oztaner, S.M.: A Word Grammar of Turkish with Morphophonemic Rules. Master's thesis, Middle East Technical University (1996)
5. Sedláček, R., Smrž, P.: A New Czech Morphological Analyser `ajka`. In: Proceedings of TSD 2001, Berlin, Springer-Verlag (2001) 100–107
6. Knuth, D.E.: The Art of Computer Programming: Fundamental Algorithms. 2nd edn. Volume 1. Addison Wesley (1973)
7. Knuth, D.E.: The Art of Computer Programming: Sorting and Searching. 2nd edn. Volume 3. Addison Wesley (1973)
8. Jackson, P.: Introduction to Expert Systems. 3rd edn. Addison Wesley Longman, Harlow, England (1999)
9. Komárek, M.: Mluvnice češtiny II (Grammar of Czech). Academia, Praha (1986) In Czech.
10. Osolsobě, K.: Algorithmic Description of Czech Formal Morphology and Czech Machine Dictionary. PhD thesis, Faculty of Arts, Masaryk University Brno (1996) In Czech.