Performing Adaptive Morphological Analysis Using Internet Resources*

Marek Trabalka and Mária Bieliková

Department of Computer Science and Engineering
Slovak University of Technology
Ilkovičova 3, 812 19 Bratislava, Slovakia
trabalka@decef.elf.stuba.sk, bielik@elf.stuba.sk

Abstract. In this paper, we describe an approach to an adaptive morphological analysis based on lexicon corpus acquired from Internet. We focus on automating categorization words into a morphological paradigm in flexive languages. It is done by inducing possible word forms using morphological knowledge base and by looking for word forms of possible inflections in a morphological lexicon.
We developed a prototype system based on the proposed approach. Our system is general (it respects language but it performs better on a flexive language). We tested the system for the Slovak language. System's lexicon is built by means of browsing Internet pages. Parsed texts, recognized to be written in Slovak, are used to establish database of Slovak words with their frequencies in texts.

1 Introduction

Most of todays research and practice in natural language processing is concerned with the English language. The concentration on English has resulted in advances in solving of its linguistic problems [5]. Morphological analysis (the process of describing a word in terms of the prefixes, suffixes, and root forms that comprise it) is such a phenomenon that in English has a little importance due to morphological simplicity of the English language. Limited amount of morphological information is often used with part-of-speech disambiguation techniques (taggers) [2, 6, 1]. By associating each word with a unique tag, tagging helps in disambiguating words with multiple parts of speech, hence tagging can provide information on ways in which each word is used.

Many other languages use more of morphology than English (e.g., German language; Slavic languages such as the Slovak, Czech, or Russian languages). Words in these languages have usually several different morphological forms that are created by changing a suffix. Each form represents some of morphological attributes like case, number or gender. Such languages with a rich morphology are called flexive.

* The work reported here was partially supported by Slovak Science Grant Agency, grant No. G1/4289/97.

In flexive languages morphology strongly influences syntax. In English sentence, meaning is determined by a word order while e.g., in Slovak it depends primarily on a word form and order of words is not so strict. On the other hand syntactical analysis of the sentence in flexive language is after successful morphological analysis simplier task than in English.

Bellow we illustrate richness of flexive languages morphology. In an example sentence below you can see two different forms of Slovak word *ruka* (hand):

In Slovak: *Ruka* drží *ruku.*
In English: *Hand* holds *hand.*

A word form (*ruka* versus *ruku*) determines its role in the sentence. In this case, *ruka* is the subject and *ruku* is the object. Following example illustrates a variability of the order of words:

Alternative 1 in Slovak: *Ruka* drží *nohu.*
Alternative 2 in Slovak: *Nohu* drží *ruka.*
In English: *Hand* holds *foot.*

where both alternatives in Slovak have the same meaning. Again we can see that determination of the word form is necessary for sentence parsing and understanding.

2 Method of Adaptive Morphological Analysis

Usually, dealing with morpohology is to define a set of paradigms expressing type of inflection and to create a morphological lexicon containing as many words as possible with reference to their appropriate paradigms [4].

Creating a morphological lexicon is a big investment in time and money. Moreover, it is an endless task. If we omit a problem with manual setting word-paradigm pairs there still exists problem of special terms and new words that should be added. General purpose applications (e.g., text editors, using the lexicon for spell-checking) usually only allow user to add a word on his own.

To support an analysis of unknown words (morphology learning) we propose a method of adaptive morphological analysis. The idea is as follows:

> *We give to computer multitude of word forms and paradigm definitions and the program should determine an appropriate paradigm for each word.*

The method uses paradigms and heuristics to compute an acceptance probability of a particular word. The acceptance probability expresses a degree of belief in the correct determination of paradigm for the word.

Input to the method is a word, which is subject to the morphological analysis. Output is determination of a form for this word (a morphological tag) with an acceptance probability and acceptance probabilities for derived morphological tag alternatives of the analysed word. These alternatives and their acceptance probabilities are computed by means of linguistic knowledge. Linguistic knowledge is either filled in advance by an expert, or learned by the system during the previous analysis.

The method consists of the following steps:

1. Analysis of possible forms that the analysed word can match. Set of form/paradigm pairs is composed.
2. Performing inflection for each alternative. Set of groups, one group of words for each form/paradigm pair is composed. Each group contains *all* word forms that can be inflected using the given paradigm.
3. Looking for all word forms created in step 2 in a lexicon. Each word form found in the lexicon is marked-up.
4. Computing the acceptance probability for each form/paradigm alternative. It strongly depends on the number of forms found in the lexicon and on heuristics applicable for a particular word form. All these partial acceptance probabilities of word forms of each group are composed to resulting acceptance probability of form/paradigm pair.

In order to perform an analysis, four different knowledge types are used:

- *Attributes* represent morphological attributes such as gender (musculine, feminine or neuter) or number (singular or plural).
- *Forms* express morphological forms. They are defined by specifying values for some subset of attributes. Comparing with English tag-sets [6] the attribute is similar to a simple tag and a form represents typical group of tags used together.
- *Paradigms* are sets of forms and relations between them. Each relation describes how to create one word form from another (e.g., by adding a specified suffix).
- *Heuristics* are sets of quantitative information that are used to express an acceptance probability of forms and paradigms.

The method of adaptive morphological analysis usually creates several morphologically non-correct forms deduced from selected paradigms alternatives. Computed acceptance probabilities distinguish these non-correct forms from the correct ones. For example in the process of analysing the word *otcom* (form of the word *otec*, which means father) 19 different forms of the analysed word are inflected: *otec, otcovia, otcov, otca, otcom, otcovi, otcoch, otcami, otc, otcu, otcmi, otcomy, otcomov, otcoma, otcomom, otcomu, otcomoch, otcome, otcommi.*

Only first 8 of them are morphologically correct Slovak words (the analysed word *otcom* is among them).

3 Acquiring Vocabulary from Internet

If computation of acceptance probabilities is based solely on knowledge base (attributes, forms, paradigms and heuristics) distance from acceptance probabilities of correct forms to acceptance probabilities of incorrect forms is sometimes very close. This distance can be increased by using a lexicon. The more (correct) forms are found in the lexicon, the bigger is this distance.

Vast repositories of useful textual information in machine-readable form are now commonly available, especially with advantage of Internet and its World Wide Web service (WWW). We have on mind that the Internet contains generally not the quality but quantity of texts. Nevertheless we used it as our primary source for creating a language corpus due to its availability.

Typical problem of electronic documents in the Slovak language (and many other languages including almost all Slavic languages) on the Internet is diacritic. Texts contain letters that standard Latin-1 charset does not contain. We should take into account this fact when design the process of text selection for analysis. Our approach is based on statistical research of the Slovak language [3]. The proportion of letters with diacritical signs is an important criterion for considering text suitable for learning (i.e., it is Slovak text).

Building of a lexicon is automated by a multiagent system that browses WWW pages and extracts words from the texts. System consists of a global synchronization server and one or more local agents. The global server functions as a centralized storage the local agents are synchronized with. The system also offers an interface for maintaining the global lexicon.

Local agents can run on a variable number of machines. They work separately and synchronize their contents with a global server. Each agent maintains its own set of hyperlinks and a private lexicon.

The process of HTML page analysis is as follows. Primary objective is to decide whether the page is written in the Slovak language. At first the HTML page is downloaded and character set is determined, it must be dedicated to Eastern Europe. Contents is re-coded into Win-1250 charset and amount of letters with diacritical signs is computed. A common Slovak text contains specific subset of charset and some minimal level of diacritics. Hence texts written without diacritics are refused. Note that nowadays a lot of Slovak texts are published on Internet without diacritic. Decision about considering text for further processing is based also on proportion of specific Slovak and non-Slovak words. Maintainer of a global lexicon adds information of whether word is specific Slovak or non-Slovak manually.

The selected text is used both to extract URL references and to create list of words that the page contains. All hyperlinks found in HTML document are added to the database of links with computed probability that it contains Slovak text.

Synchronization with a global server is done in both directions. Agent sends words with their occurrences to the server and the server sends back its complete marked-up lexicon. Agent remembers last synchronization of each word so it sends only occurrences added after the last synchronization.

Final result of such a system is a collection of words forms together with number of their occurrences in texts.

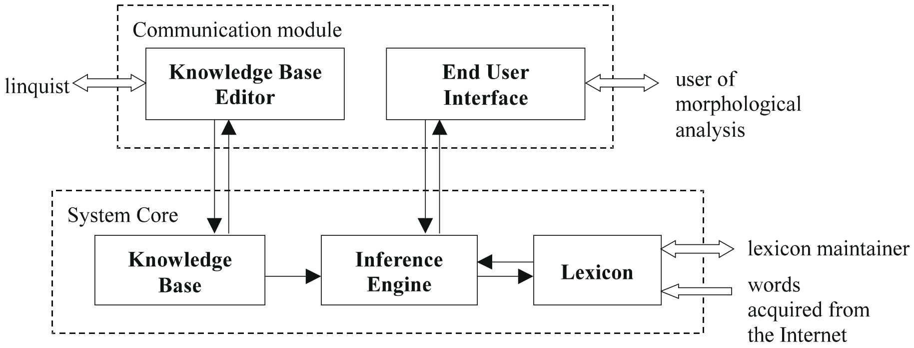

Fig. 1. Architecture of the system for morphological analysis.

4 Experiments

We developed a system that is based on the proposed method of adaptive morphological analysis. The overall structure of the system is depicted on Figure 1. In order to support creation and modification of morphological knowledge we developed also an editor of the morphological knowledge base. It is used especially by language experts. The knowledge base editor allows to insert or modify paradigm and form definitions together with heuristics used by the inference engine to perform morphological analysis.

The system core implements our method of adaptive morphological analysis. It is implemented as a series of programs in interpreted PERL. The system offers functions for analysis and synthesis of morphological forms. The result of analysis is:

– possible forms that match the input word and also
– acceptance probability of each case.

Co-operation of the two developed systems (a system for morphological analysis and system for automation of building a lexicon) is important. Our adaptive morphological analysis obviously needs the lexicon of morphologically correct words. But also the system for acquiring the lexicon can use morphology to improve its decision process by inflection of non-marked words and searching for their other forms.

We developed a useful tool that uses both of these systems. We made a special World Wide Web site containing searching engine interface where user can search for all or selected word forms. The user simply puts a word and selects the searching engine. Our system recognizes word paradigm, creates all word forms and sends them to desired searching engine, e.g. Altavista.

Although system is still under development it already offers some usable results. System for automation of building a lexicon started with the initial lexicon containing about 800 words only. After approximately 100 hours of running, the agents processed more than 25.000 links. Only 510 of them successfully passed tests and were reporting to be regular Slovak texts with diacritics. More than 200.000 of words were found on these pages that finally represent 50.000 unique word forms. We manually checked most frequent words (around 20.000). The analysis showed that none of the words that occurred in more than 5 different pages were incorrect.

After testing output quality of acquired lexicon we began to use it for testing proposed adaptive morphology concept. We selected manually from acquired lexicon words marked as Slovak and used it for testing morphology based on a set of paradigms for Slovak language [4]. We tested for example influence of multiple occurrences word forms on the final result of morphological analysis. The lexicon was modified by removing some word forms to recognize changes it was causing.

By increasing the number of forms in the lexicon we observed an increase of the acceptance probability of the input word form. Correct forms had usually doubled acceptance probability. But if the only few word forms are in the lexicon, result success strongly depends on which word form is in the lexicon.

5 Conclusion

In this paper, we describe a method of adaptive morphological analysis based on usage of real texts as a source of word forms. We developed a software system that implements proposed method and scans Internet for texts used in process of lexicon construction. Searching interface is currently under development.

The advantage of the developed system is that it also works well for unknown words (not included in the lexicon yet). The system recognizes possible paradigms and creates a list containing all possible forms.

The developed system can be primarily used to enhance full-text searching regardless of form of searched word. It also sets up a basis for a syntactic analysis, which we are currently working on.

References

1. Brill, E.: A simple rule-based part of speech tagger. In Proceedings of the Third Annual Conference on Applied Natural Language Processing. 1992.
2. Chandrasekar, R., Srinivas, B.: Using syntactic information in document filtering: A comparative study of part-of-speech tagging and supertagging. In Proceedings of RIAO97, Montreal, pages 531-545. 1997.
3. Mistrík, J.: Frequency of forms and constructs in Slovak. Bratislava, Veda. 1985.
4. Páleš, E.: Sapfo: Natural language paraphraser for Slovak language. Bratislava. 1993.
5. Shinghal, R.: Formal concepts in AI: Fundamentals. Chapter 6. Natural Language Processing: A prescriptive grammar. Chapman & Hall, London. 1992.
6. van Guilder, L.: Automated part of speech tagging: a brief overview. Georgetown University. 1995.