A Method for Development of Automatic Morphological Analysis Systems for Inflective Languages*

Alexander Gelbukh and Grigori Sidorov

Natural Language Laboratory,
Center for Computing Research (CIC),
National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz, CP 07738, Zacatenco, México DF
{gelbukh,sidorov}@cic.ipn.mx
www.gelbukh.com, cic.ipn.mx/~sidorov

Abstract. A crucial problem in development of the systems of automatic morphological analysis for inflective languages is treatment of stem alternations. The existing models require the development of the corresponding rules that specify which stems can be generated from the given one. A large number of such rules (e.g., for Russian about 1000) do not have any reasonable linguistic interpretation. We suggest a method that avoids the use of such rules by generating the hypotheses about possible grammatical forms, storing all possible stems in the dictionary, and verifying the hypotheses using the generation module. Methods of such type are known as "analysis through generation"; they make the system development much simpler than the standard direct approach. A system of morphological analysis and synthesis for Russian language developed with this method is available for free for academic use; see www.gelbukh.com, cic.ipn.mx/~sidorov.

1 Introduction

There exist many different models of the automatic morphological analysis for different languages. One of the most famous models is the two-level model (KIMMO) suggested by Kimmo Koskenniemi [4]; there are many other models [1, 3, 5, 7, 8].

Nevertheless, different languages have different morphological structure and the methods that are perfectly suitable for languages with poor morphology (like English) or with agglutinative morphology (like Finnish) are not the best methods for inflective languages (like Spanish, Czech, or Russian). By "not the best method" we mean that, since the morphological system of any inflective language is finite, any method of analysis gives correct results, but not all methods are equally convenient and easy to implement.

An extreme point is storing all grammatical forms in a dictionary (such a method is known as "bag of words"), which is possible for inflective languages, though

* Work done under partial support of CONACyT, SNI, and CGEPI-IPN, Mexico.

impossible for agglutinative or polysynthetic ones. Modern computers have the possibility of storing large databases containing all grammatical forms for inflective languages (a rough approximation for Spanish and Russian is 20 to 50 megabytes). Still applications that use more sophisticated algorithms to reduce the dictionary size to, say, 1 megabyte, are preferable. Indeed, a morphological analyzer is usually used together with a syntactic parser, semantic analyzer and a reasoning or retrieval engine, and freeing physical memory for these modules is highly desirable (of course, the use of large virtual memory makes simultaneous access to very large data structures possible, but does not make it faster).

Algorithmic (non-"bag-of-words") solutions have a number of additional advantages. For example, such algorithms have the possibility to recognize unknown (new) words. This is a crucial feature for a morphological analyzer since new words constantly appear in the language, not speaking of that no dictionary can be complete.

In this paper we discuss how to develop an automatic morphological analysis system for an inflective language spending less effort and applying more intuitive and flexible morphological models. We show that the use of not so straightforward method can greatly simplify the analysis procedure and allows using morphological models much more similar to the traditional grammars.

The main idea of the method is to avoid development of stem transformation rules in analysis and to use instead the generation module (this idea is known in Artificial Intelligence as "analysis through generation"). Implementation of this idea requires storing in the morphological dictionary all stems for each word with the corresponding information.

In the rest of the paper, we first describe the suggested method (types of information, types of morphological models, etc.) and then briefly discuss its implementation for Russian language (the corresponding system is freely available for academic use).

2 The Method

As we have mentioned, the main problem of automatic morphological analysis of inflective languages is usually stem alternations. The direct way to resolve this problem is constructing the rules that take into account all possible stem alternations during the analysis process; for example, for Russian the number of such rules is about a thousand [5]. However, such rules do not have any correspondence in traditional grammars; in addition, they have no intuitive correspondence in language knowledge; finally, they are too many.

Another possibility is to store all stems in the dictionary, together with the information on their possible grammatical categories; this method was used for Russian [2] and for Czech [6]. We adopt this possibility, but propose a different technique for treatment of grammatical information: our technique is dynamic while the techniques described in [2, 6] are static.

Namely, we apply the technique known in Artificial Intelligence as "analysis through generation." Since analysis is usually more complex than generation, this technique allows for simpler implementation.

In the next subsections, we describe the types of morphological information we use; then we discuss the morphological models (and the corresponding algorithms) we have used to implement the method; and finally, we describe the functioning of our method: analysis, generation, and treatment of unknown (new) words.

2.1 Types of Grammatical Information

We use two types of information:

- the stem dictionary and
- a list of grammatical categories.

The information about the stems is stored in the morphological dictionary. This information is basically the data needed for generation, such as:

- Part of speech,
- The presence of alternations,
- Grammatical type (in Russian, there are three genders and for each gender there are several word formation types: say, for feminine there are 7 types, etc.),
- Special marks: for example, in Russian some nouns have two forms of the prepositional case (*шкафу* versus *шкафе* '(in) wardrobe' versus '(about) wardrobe'), which should be marked in the dictionary.

We explicitly store in the dictionary all variants of stems as independent forms, together with a stem number (first stem, second stem, etc.). In Russian, nouns with alternations have two possible stems and verbs have up to four stems.

Anther type of information is a list of grammatical categories; they are represented as sets of grammemes. By grammeme we mean a value of a grammatical category, such as "singular" (one of the values of the category "number") or "nominative case" (one of the values of the category "case"). Thus, any grammatical form is characterized by a set of grammemes. For example, for a Russian noun this set contains a value of case and of number; for a Russian full adjective it is a value of case, number, and gender.

2.2 Types of Morphological Models

Three morphological models are used:

- The correspondence between the flections and the grammemes,
- The correspondence between the stems and the grammemes,
- The correspondence between alternating stems of the same lexem.

The first model establishes the correspondence between the flexions and sets of grammemes, taking into account different grammatical types fixed in the dictionary. In the process of analysis, we use the correspondence "flexions $\Rightarrow$ sets of grammemes" (it is used to formulate hypothesis), and in the process of generation, the correspondence "sets of grammemes $\Rightarrow$ flexions".

A similar correspondence is established between the sets of grammemes and the types of stems; however, this correspondence is used only for generation. For example, if a Russian masculine noun of a certain type has a stem alternation, then the first stem is used for all forms except for genitive (case) plural, for which the second stem is used. Note that a corresponding model for analysis is unnecessary, which makes our method simpler than direct analysis.

To be able to generate all forms starting from a given form, it is necessary to be able to obtain all variants of stems from the given stem. There are two ways to do this: static and dynamic, which have their own pros and contras. The static method implies storing in the dictionary together with the stems the correspondence between them (e.g., each stem has a unique identifier by which stems are linked within the dictionary). Storing the explicit links increases the size of the dictionary.

Thus, we propose to do this dynamically. Namely, the algorithm of constructing all stems from a given stem is to be implemented (in fact, it must be implemented anyway since it is used to compile the dictionary). It is sufficient to develop the algorithm for constructing the first stem (that of the dictionary form, such as infinitive) from any other stem, and any other stem from the first stem. In this way, starting from any stem we can generate any other stem. The difference between static and dynamic method is that in the former case, the algorithm is applied in the compile time (when the dictionary is built) while in the latter case, in runtime.

Note that the rules of these algorithms are different from the rules that have to be developed to implement analysis directly. For Russian, we use about 50 rules, intuitively so clear that in fact any person learning Russian learns (and understands) these rules of stem construction. Here is an example of a stem transformation rule:

$$\text{-VC, } * \Rightarrow \text{(V)} \qquad (1)$$

which means: if the stem ends in a vowel (V) following by a consonant (C) and the stem type contains the symbol "*" (we use the types from [9]) then remove this vowel. Being applied to the first stem of the noun *молоток* 'hammer', the rule gives the stem *молотк-(а)* 'of hammer'.

This contrasts with about 1000 rules necessary for direct analysis, which in addition are very superficial and anti-intuitive. For example, to analyze a non-first-stem word, [5] uses rules that try to invert the effect of (1): if the stem ends in a consonant, try to insert a vowel before it and look up each resulting hypothetical stem in the dictionary (for *молотк-(а)*, try *молотек-*, *молоток-*, etc.). This also affects the system performance.

Two considerations explain the simplicity of our rules. First, we use the information about the type of the stem stored in the dictionary. For Russian, this information is available from the dictionary [9]. Second, often generation of a non-first stem from the first one is simpler than vice versa. More precisely, the stem that appears in the dictionaries for a given language is the one that allows simpler generation of other stems (note that in some languages for, say, verbs this is not the infinitive: say, in Hebrew this is 3rd person past singular masculine).

2.3 Data Preparation

Our method needs some preliminary work of data preparation, carrying out the following main steps:

- To describe and classify all words of the given language into unique grammatical classes;
- To convert the information on the words into a stem dictionary (generating only the first stem);
- To apply the algorithms of stem generation (from the first stem to other stems) to generate all stems;
- To generate the special marks and the stem numbers for each (non-first) stem.

To perform the last two steps, the data record generated for the first stem is duplicated, the stem is changed to obtain the required non-first stem, and the mark with the stem number is added.

2.4 Generation Process

The generation process is simple. Given the data from the dictionary (including the stem and its number) and a set of grammemes, it is required to build a wordform of the same lexeme that would have the given set of grammemes. Using the models we have constructed, the flexion is chosen and the necessary stem is generated (if a non-first stem was given, then we generate the first stem and from it, the necessary stem). Finally, we concatenate the stem and the flexion.

If necessary, this process is repeated several times to add more than one flexion to the stem. For example, Russian participles (which are verbal forms) have the same flexions as adjectives (which express the number and gender) and also special suffixes (which indicate that this is a participle), i.e., they are a concatenation of a stem and two affixes: a suffix and a flexion (*пис-ать* → *пиш-ущ-ий* 'writ-e' → 'writt-en'). In this case, we first generate the stem of participle by adding the suffix (we use the information from the dictionary on the properties of the corresponding verbal stem) and then change the dictionary information to the information for an adjective of the corresponding type. In case of Russian, such recursion is limited to only three levels (one more level is added due to the reflexive verbs that have a postfix morpheme *-ся*, similar in function to the Spanish *-se*: *пиш-ущ-ий-ся* 'is written').

2.5 Analysis Process

Given an input letter string (a wordform), we analyze it in the following way:

1. The letters are separated one by one from right to left to get the possible flexion (the zero flexion is tried at the first step): given *stopping*, we try *-∅* (zero flection); at the next iterations *-g*, *-ng*, *-ing*, *-ping*, etc. are tried.

2. If the flexion (here *-ing*) is found in the list, we apply the model "flexions ⇒ sets of grammemes", which gives us a hypothesis about the possible set of grammemes (here "verb, present participle").
3. Then we obtain the information for the rest of the form (the potential stem, here *stopp-*) from the dictionary.
4. Finally, we generate the corresponding grammatical form according to our hypothesis and the dictionary information (here, the generated past participle of the verbal stem *stopp-* is *stopping*).
5. If the obtained result coincides with the input form, then the hypothesis is accepted. Otherwise, the process repeats from the step 1.

If the wordform consists of several morphemes (a stem and several affixes), then the analysis process is recursive, precisely as generation. In case of Russian, only 3 levels of recursion are sufficient.

As one can see, our method of analysis is not more complex than generation. The only added modules are the model "flexions ⇒ sets of grammemes" and the module of interaction between different models.

2.6 Treatment of Unknown Words

The treatment of unknown words is also simple. We apply the same procedure of analysis to single out the hypothetical stem. If the stem is not found in the dictionary, we use the longest match stem (matching the strings from right to left) compatible with the given set of affixes. The longest match stem is the stem present in the dictionary that has as long as possible *ending* substring in common with the given stem (and is compatible with already separated affixes).

In this way, for example, an (unknown) input string *sortifies* will be analyzed as *classifies*: verb, 3rd person, present, singular, given that *classifi-* is its longest match stem for *sortifi-* (matching by *-ifi-*) compatible with the affix *-es*.

To facilitate this search, the stem dictionary is ordered by inverse order, i.e., the stems are ordered lexicographically from right to left (by the last letter; if they coincide then last but one, etc.).

Note that the systems like [1, 2] based on the left-to-right order of analysis (first separating the stem and only then analyzing the resting affixes) have to imitate this process with a special dictionary of, say, a list of 5-letter stem endings, since in such systems the main stem dictionary is ordered by direct order (left to right, by first letters).

3 Implementation for Russian

We have implemented this method for Russian [7], which is a highly inflective language. A morphological dictionary [9] including about 100,000 lexemes was used. Fortunately, the models for Russian morphology had been developed in the same dictionary. This dictionary is oriented for generation: using it, a person without any knowledge of Russian can generate any given Russian morphological form. Due to

the used technique, it proved to be possible to borrow all grammatical types from this dictionary without changing.

The implementation process took less than one year of work of one person. The system for Russian is available for free for academic use as Windows DLL or EXE file.

4 Conclusions

We have presented a method for building systems of automatic morphological analysis for inflective languages. The method is based on storing of all variants of the stems in the dictionary and on application of the method known as "analysis through generation." This gives a simple way to develop morphological systems for inflective languages.

The method has been implemented for Russian language with minimum resources in a rather little time period. The system for Russian is freely available for academic use as Windows DLL or EXE file; for licensing write at the addresses indicated at the title page of this paper.

References

1. Gelbukh, A. A data structure for prefix search under access locality requirements and its application to spelling correction. Proc. of *MICAI-2000: Mexican International Conference on Artificial Intelligence*, Acapulco, Mexico, 2000.
2. Gelbukh, A. An effectively implementable model of morphology of an inflective language (in Russian). *J. Nauchno-Tehnicheskaya Informaciya (NTI)*, ser. 2, vol. 1, Moscow, Russia, 1992, pp. 24-31.
3. Hausser, Ronald. Three principled methods of automatic word form recognition. Proc. of *VEXTAL: Venecia per il Tratamento Automatico delle Lingue*. Venice, Italy, Sept. 1999. pp. 91-100.
4. Koskenniemi, Kimmo. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. University of Helsinki Publications, N 11, 1983.
5. Malkovsky, M. G. *Dialogue with the artificial intelligence system* (in Russian). Moscow State University, Moscow, Russia, 1985, 213 pp.
6. Sedlacek R. and P. Smrz, A new Czech morphological analyzer AJKA. Proc. of *TSD-2001*. LNCS 2166, Springer, 2001, pp 100-107.
7. Sidorov, G. O. Lemmatization in automatized system for compilation of personal style dictionaries of literature writers (in Russian). In: *Word by Dostoyevsky* (in Russian), Moscow, Russia, Russian Academy of Sciences, 1996, pp. 266-300.
8. Yablonsky, S. A. Russian morphological analysis. Proc. of *VEXTAL: Venecia per il Tratamento Automatico delle Lingue*. Venice, Italy, Sept. 1999, pp 83-90.
9. Zalizniak, A.A. *Grammatical dictionary of Russian* (in Russian). "Russkij Jazyk", Moscow, Russia, 1980.