Approach to Construction of Automatic Morphological Analysis Systems for Inflective Languages with Little Effort*

Alexander Gelbukh and Grigori Sidorov

Center for Computing Research (CIC),
National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz, esq. Miguel Othón de Mendizábal,
Mexico D. F., Zacatenco, CP 07738, Mexico
{gelbukh, sidorov}@cic.ipn.mx

Abstract. The problem of developing of systems of morphological analysis for inflective languages is known as rather laborious. We suggest the approach for development of these systems that permits to spend less time and effort. It is based on static processing of stem allomorphs and the method of analysis known as "analysis through generation". These features allows for using the morphological models oriented to generation instead of developing special models. Normally, such models are presented in traditional grammars and correspond very well to the intuition of speakers. Systems, based on this approach, were developed for Russian and Spanish with relatively little effort and time.

1 Introduction

Languages can have poor morphology (so called, analytic languages, when the grammar categories normally are expressed outside the word), like, say, English or Chinese; or rich morphology (so called, synthetic languages, when the grammar categories normally are expressed inside the word), like Finnish or Russian. At the same time, synthetic languages present two techniques of morphological arrangement:

- Fusion (when there is a tendency to express all grammatical categories by one flexion and there exist non-predictable stem alternation), like, say, in Russian, Czech and other Slavic languages, and also in Spanish or Portuguese (such languages are called **inflective**) and
- Agglutination (when there is a tendency to use different morphemes for each grammatical category and there are no stem alternations or these alternations are predictable), like in Finnish or Turk languages.

* This work was done under partial support of Mexican Government (CONACyT, SNI), National Polytechnic Institute, Mexico (CGEPI, COFAA, PIFI), and RITOS-2 net of Subprogram VII of CYTED.

In this paper, we discuss an approach that allows for rapid development with little effort (in comparison with other approaches) of systems for automatic morphological analysis/generation for inflective languages.

Morphological systems of inflective languages are finite (usually about 2-3 million grammatical word forms for a dictionary of about 100,000 words), so, finally, any method leads to the same results. Still, there are the differences in time and effort spent applying different methods. In addition, there is the difference in similarity of the used models and the models described in traditional grammars. In our opinion, the more similar these two kinds of models are, the better the system is, because computational models based on traditional grammar models are much clearer intuitively and it is much easier to apply them in the system's development.

The main problem in automatic morphological analysis of inflective languages is the treatment of non-predictable stem alternations. Indeed, if there are no such stem alternations, then the algorithm of morphological analysis is very straightforward. First, we assign a morphological class for each stem that uniquely defines a set of flexions. There is only one stem in the dictionary for each word because there are simple rules to build its allomorphs. During the morphological analysis of a wordform, we find the flexion in the wordform and, after this, the stem (the rest of the wordform, maybe, it is modified according to the rules) is searched for in the dictionary. If the flexion is compatible with the stem, then the analysis is finished. This is the case of agglutinative languages, like Finnish or Turk.

Note that our point in this paper is not to discuss the formalism that, generally speaking, can be a two-level model, direct programming, interpreter of grammar tables, or unification grammar, but the approach to treatment of stem allomorphs that does not depend on formalism.

2 Some Considerations for Inflective Languages

The case of non-predictable stem alternations is more complicated. There are two important points to discuss:

- Method of processing of stem allomorphs (static or dynamic), and
- Morphological models that are used ("artificial" models based on direct approach or "natural" models based on "analysis through generation" approach).

2.1 Static vs. Dynamic Methods

There are two methods of processing of stem allomorphs: static and dynamic (sometimes, the terms "allomorphs vs. morpheme" is used [2]). Static method means that all stem allomorphs are stored in the dictionary (normally there are 2-4 allomorphs, so the dictionary size is not significantly affected; note that normally the majority of words (say, in Russian, more than 70%) does not have stem alternations).

The allomorphs are generated beforehand, what is not difficult because the information about each stem is available.

Dynamic method means that allomorphs are constructed dynamically trying to reconstruct an allomorph that is stored in the dictionary. In inflective languages, the corresponding rules cannot be standardized, so a number of such rules is very significant (more than 1000 rules are mentioned in [4]). Besides, they do not have any intuitive correspondence in knowledge of the language. For example, in order to generate the dictionary stem for Russian *okon-* (*window*), it is necessary to delete *-o-* (*okn-*). The corresponding example for English can be (just to demonstrate the kind of processes): for *took* it is necessary to change *-oo-* to *-a-* and add *-e* to obtain *take*. It is difficult, because we do not have any beforehand information about the possible type of stem, so it is necessary to develop and apply many unintuitive rules. This method is "of high complexity (NP-complete)" ([2], p.255).

Therefore, the static method is more reasonable and easy to implement than the dynamic one for inflective languages. On the contrary, for agglutinative languages it is easy to use the dynamic method, because the rules are rather simple and intuitive. For example, the dynamic method was applied in the well-known two-level morphology ([3]) that was created for Finnish. Indeed, the idea of the two-level morphology is to create the correspondence between the abstract level of morphemes and the level of their realizations, i.e., the allomorphs (these are the two levels), using the formalism of the finite-state automata. We are not discussing formalisms in this paper, but the formalism is used in the two-level model for implementation of rules of correspondence between two levels, i.e., the dynamic processing of stem alternations. As it was mentioned before, it is possible to use this kind of processing for inflective languages, but it requires the development of substantially greater number of rules that are less intuitive in these languages.

2.2 Morphological Models

The other dilemma deals with the kind of morphological models. The obvious direct way for developing the morphological models is to create a new morphological class for any paradigm that exists in the language, thus, the number of classes is calculated up to 1500 for Czech ([5]) or 1000 for Russian ([1]). These classes are artificial, created for the purposes of analysis.

The other possibility is to use the morphological models that already exist for generation, say, in case of Russian there are about 40 morphological classes. These models are described usually in traditional grammars, because these grammars are oriented to generation. Besides, they correspond very well to the intuition of speakers. To be able to use these models, it is necessary to apply a trick that allows for applying generation instead of direct analysis. Normally, generation is much simpler than analysis. This trick is known, say, in artificial intelligence as "analysis through generation". In our case, it is applied as follows: first, the system generates all possible hypotheses based on the possible flexions, and then tries to generate the grammar forms according to each hypothesis using the corresponding stem and the stem morphological class taken from the dictionary. Note that there is a small number of classes, but the peculiarities of words are described using grammar marks for

words, like, for example, the presence of alternations or the absence of singular (pluralia tantum), etc. These marks are interpreted during the process of analysis/generation.

Obviously, it is much easier for development of a system to have a small number of morphological classes, which correspond very well to the intuition of speakers. Sometimes these classes already exist, but if not, it is easier to characterize the words in a given language applying the simple and intuitive classification.

We suggest using during analysis the models created for generation, but there is the other possibility to apply the same models. It is possible to generate all possible wordforms beforehand and the process of analysis is just search in the database of these forms (bag of wordforms). This is the other possibility to apply analysis through generation. Its advantage is simplicity of the algorithm of analysis: it is just a search, but note, that anyway an additional algorithm should be developed for generation of all forms. Still, its disadvantage is the size of the dictionary that is much more than the size of the dictionary of stems. The exact number depends on the number of grammar forms per lemma in a language, e.g., in Russian it is more than 30 times. Thus, there is the choice: large dictionary and very simple algorithm vs. small dictionary and more sophisticated algorithm. From this point of view, the algorithm of analysis can be considered as a method of compression of the dictionary (and a very good compression, indeed). There are other advantages of usage of the algorithm of analysis over the bag of wordforms because the algorithm possesses additional grammar knowledge. For example, processing of ungrammatical forms like **taked*, when the algorithm understands what is meant and suggests the correct form *took*.

3 Approach

We suggest using static method of processing of stem allomorphs, i.e., all allomorphs are stored in the dictionary, and applying the natural morphological models created for generation based on "analysis through generation" procedure.

The first stage is data preparation. The words of a language should be characterized in terms of the used morphological models. Then the stem dictionary is generated with all possible allomorphs of each stem. Note that the stem allomorphs should be marked according to the algorithm of their generation, for example, first, second, etc. This information is necessary during grammar form generation, namely, for choosing the correct stem allomorph.

The next stage is the development of the algorithm of morphological analysis. The following modules (parts of algorithm) are necessary:

- Module of generation of hypotheses (correspondence between flexions and sets of possible values of grammar categories (flexion → values), e.g., in English, flexion *–s* can express plural for nouns or 3 person singular for verbs, etc.).
- Module of choice of stem allomorphs (correspondence between sets of values of grammar categories of morphological classes and number of stem allomorph (values → number of a stem allomorph), e.g., in English, if we consider the verb stems *verify/verifi-* as allomorphs, then the first allomorph

is used for the present (not 3rd person, singular), and the second one for the past or present 3rd person singular, etc.); this can be done using masks, patterns, direct programming, etc. Note that we do not need the reverse correspondence because we apply this module only in generation.

- Module of choice of flexions (which flexion is used for a given set of grammar categories of a given class (values → flexion), e.g., in English, for plural of nouns the flexions *–s* or *–es* are used depending on the stem last letters).
- Module of processing of irregular forms. All irregular grammar forms (like irregular verbs, etc.) are stored in the dictionary with their lemma and values of grammar categories (number, tense, etc.). Therefore, their analysis is just searching in the dictionary (we should always have hypothesis of the irregular form with zero flexion). Their generation is also searching in the dictionary, but for the lemma and the corresponding values of grammar categories.

The procedure of generation is very simple. The input is a set of values of grammar categories and a string that identifies the word (stem allomorph or lemma). The procedure implies obtaining the information from the dictionary (the morphological class, etc.), choice of the correct stem allomorph, and choice of the correct flexion (see corresponding modules).

The procedure of analysis also is not complicated. The input is a string of characters. The procedure is as follows:

- Remove character by character from the string in order to find the possible flexion and the stem (also zero flexion is always considered),
- Formulate the hypotheses for the flexion,
- Call the generation procedure for each hypothesis,
- Compare the result of generation with the input. If they coincide then the hypothesis is correct.

Note that it is important to apply generation because otherwise there would be incorrect forms that are analyzed as the correct ones, for example, **taked* (instead of *took*), because both the stem *take-* and the flexion *-d* exist, but they are incompatible, that is verified through generation (the correct form *took* will be generated and the forms do not coincide).

If there are several affixes in a word (for example, in Russian there are suffixes of participles), then this procedure of analysis can be applied recursively. This situation is not typical for inflective languages. In this case, the important action in the algorithm is to change the grammar information obtained from the dictionary to the grammar information that corresponds to these affixes (for example, these participles in Russian have verbal stem but they have the same morphological class as adjectives).

4 Conclusions

We presented the approach for developing systems of morphological analysis for inflective languages. Our approach allows for spending less time and effort for this development. The approach is based on the static method of processing of stem allomorphs and the procedure of analysis that is known as "analysis through generation". These features allows for the usage of the morphological models that are oriented for generation. These models are much more simple and intuitive than the specially developed models for analysis. Frequently these models can be taken from the traditional grammars.

We applied this approach for the development of the systems of morphological analysis for Russian and Spanish with sufficiently large dictionaries (100,000 word and 40,000 words correspondingly). The process of development was relatively simple and fast even for such language with rather complicated morphology as Russian. It took about 6 months of development by one person for Russian (it could have been less if the approach would be accepted beforehand), and it took about 2 months of development for Spanish by one master student. The dictionaries oriented to generation existed in both cases. It took some part of the mentioned time to prepare them (transform to database format and generate the stem allomorphs). These systems are freely available for academic use.

References

[1] Gel'bukh, A.F. Effective implementation of morphology model for an inflectional natural language. *J. Automatic Documentation and Mathematical Linguistics*, Allerton Press, vol. 26, N 1, 1992, pp. 22–31.

[2] Hausser, Roland. *Foundations of Computational linguistics*. Springer, 1999, 534 p.

[3] Koskenniemi, Kimmo. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. University of Helsinki Publications, N 11, 1983.

[4] Malkovsky, M. G. *Dialogue with an artificial intelligence system* (in Russian). Moscow State University, Moscow, Russia, 1985, 213 pp.

[5] Sedlacek R. and P. Smrz, A new Czech morphological analyzer AJKA. Proc. of *TSD-2001*. LNCS 2166, Springer, 2001, pp. 100–107.

[6] 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.

[7] Sproat, R. *Morphology and computation*. Cambridge, MA, MIT Press, 1992, 313 p.