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, www.gelbukh.com

Abstract. Development of morphological analysis systems for inflective languages is a tedious and laborious task. We suggest an approach for development of such 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 allow for using the morphological models oriented to generation, instead of developing special analysis models. Normally, generation 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.

1 Introduction

Languages can have poor morphology (so called analytic languages, in which the grammar categories normally are expressed by standalone functional words), e.g, English or Chinese, or rich morphology (so called synthetic languages, in which the grammar categories normally are expressed inside the word), e.g., Finnish or Russian. Synthetic languages can use one of the following two ways of morphological arrangement:

- *Agglutination* (a tendency to use a separate morpheme for each grammatical category; there are no stem alternations or these alternations are predictable), e.g., Finnish or Turk languages;
- *Fusion* (a tendency to express all grammatical categories by one flexion; often implies complex non-predictable stem alternations), e.g., Russian, Czech, and other Slavic languages, Spanish, Portuguese; such languages are called **inflective**.

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.

[*] Work was done under partial support of Mexican Government (CONACyT, SNI), IPN, Mexico (CGEPI, COFAA, PIFI), and RITOS-2.

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

Note that our point in this paper is not to discuss the formalism (different formalisms can be used with our method) but the approach to the treatment of stem allomorphs, which does not depend on formalism.

2 Some Considerations on Inflective Languages

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: (1) Beforehand, we assign a morphological class to each stem that uniquely defines a set of flexions; it is enough to store only one stem in the dictionary for each word because there are simple rules to build all its allomorphs. (2) During the morphological analysis of a given wordform, we find the flexion in the wordform and after this, the stem (the rest of the wordform; it may be modified according to the rules) is looked up in the dictionary. (3) If the flexion is compatible with the stem, then the analysis is finished. This is the case of agglutinative languages, like Finnish or Turk.

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

- Method of processing of stem allomorphs (static versus dynamic), and
- Morphological models used ("artificial" models for the direct analysis approach versus "natural" models for the "analysis through generation" approach).

2.1 Static vs. Dynamic Methods

There are two methods of processing of stem allomorphs (sometimes, the terms "allomorphs vs. morpheme" is used [2]): static and dynamic. 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 any stem alternations). The allomorphs are generated beforehand, which is not difficult because the information about each stem is available.

Dynamic method means that the allomorphs are constructed dynamically basing on only one dictionary record. In inflective languages, the corresponding rules cannot be standardized, so the number of such rules is very large (more than 1000 rules are mentioned in [4]). Besides, they do not have any intuitive correspondence in common 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 (in English, there are much fewer such words): 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 computational 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 initially developed 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). It is dynamic processing of stem alternations that is used the two-level model for implementation of rules of correspondence between the two levels. As we have mentioned before, for inflective languages it *is* possible to use this kind of processing, however, this requires development of much greater number of rules, which are less intuitive in these languages.

2.2 Morphological Models

Another choice 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; with this, 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 as few as about 40 morphological classes. These models are usually described 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 method that allows for applying generation instead of direct analysis, because usually generation is much simpler than analysis. This method is known 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 its morphological class taken from the dictionary. Note that the number of classes is small, while the peculiarities of words are described using morphological marks stored in the dictionary entries for specific words, for example, the presence of alternations, 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 to use during analysis the models created for generation. However, there is another possibility to apply the same models. Namely, it is possible to generate all possible wordforms beforehand and during the process of analysis just to search in the database that stores all these forms. This is another possibility to apply analysis through generation. Its advantage is the simplicity of the analysis algorithm: it is just a lookup (note, however, that in any case an additional algorithm is to be developed for generation of all forms). Still, its disadvantage is the size of the dictionary that is much greater than that of the dictionary of stems. The exact number depends on the number of wordforms per lemma in a language, e.g., in Russian it is more than 30 times. Thus, the choice is: a large dictionary and a very simple algorithm versus small dictionary and more sophisticated algorithm. The latter can be viewed as compression of the dictionary (and a very good compression, indeed). There are other advantages of using of the algorithm of analysis over the database of wordforms: the algorithm possesses additional grammar knowledge. For example, processing of ungrammatical forms like **taked*, the algorithm can understand what is meant and suggest the correct form *took*.

3 Approach

We suggest using static method of processing of stem allomorphs (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 morphological models used. Then, the stem dictionary is generated with all possible allomorphs of each stem. Note that the stem allomorphs should be marked according to the rule of their generation, for example, first, second, etc. stem. This information is necessary during wordform 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 the algorithm) are necessary:

- Module of hypothesis generation (the correspondence between the flexions and the sets of possible values of grammar categories (flexion • values), e.g., in English, flexion *–s* can express plural for nouns or 3rd person singular for verbs, etc.).
- Module of choice of stem allomorphs (the correspondence between the sets of values of grammar categories of morphological classes and the number of the stem allomorph (values • stem allomorph number), e.g., in English, if we consider the verb stems *verify/verifi-* as allomorphs, then the first allomorph is used for the present tense (except for 3rd person, singular) and the second one for the past tense or present tense 3rd person singular, etc. This can be done using bit patterns, direct programming, etc. Note that we do not need the inverse 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 the noun flexions *–s* or *–es* are used depending on the stem's final letters.
- Module of processing of irregular forms. All irregular grammar forms (irregular verbs, etc.) are stored in the dictionary with their lemma and the values of grammar categories (number, tense, etc.). Therefore, their analysis consists just in looking up in the dictionary (we should always check the hypothesis of the irregular form with zero flexion). Their generation is also consists in looking up in the dictionary—for the lemma and the corresponding values of grammar categories.

The generation procedure is very simple. Its input is a set of values of grammar categories and a string that identifies the word (stem allomorph or lemma). The procedure implies (1) obtaining the information from the dictionary (the morphological class, etc.), (2) choosing the correct stem allomorph, and (3) choosing the correct flexion (see the corresponding modules).

The analysis procedure also is not complicated. Its input is a character string. The procedure works as follows:

- Remove characters from the string one by one in order to find the possible flexion and the corresponding stem (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 some incorrect forms would be accepted by the analyzer (overgeneration), for example: for **taked* (instead of *took*), both the stem *take-* and the flexion *-d* (past tense) do exist, but they are incompatible, which is verified through generation (the correct form *took* for past tense 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 analysis procedure can be applied recursively. (This situation, however, is not typical for inflective languages.) In this case, the algorithm is to change the grammar information obtained from the dictionary to the grammar information that corresponds to these affixes (for example, those Russian participles have verbal stem but they have the same morphological class as adjectives).

4 Conclusions

We have presented the approach for developing systems of morphological analysis for inflective languages. It allows for spending less time and effort on this development. It is based on the static method of processing of stem allomorphs and the procedure of analysis known as “analysis through generation.” These features allow using the morphological models that are oriented to generation. These models are much simpler and much more intuitive than those specially developed for analysis. Frequently these models can be taken from the traditional grammars.

We have applied this approach for the development of the systems of morphological analysis for Russian and Spanish with sufficiently large dictionaries (100,000 and 40,000 words, correspondingly). The development process was relatively simple and fast even for such morphologically rich language as Russian: it took about 6 months of development by one person for Russian (it would take less if we follow this approach from the very beginning) and about 2 months for Spanish of one master student. In both cases, the dictionaries oriented to generation were already available. Some part of the time was devoted to their preparation (transformation to database format and generation of the stem allomorphs). These systems are freely available for academic (see the author’s contact address).

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.