INTERNATIONAL CONFERENCE ON
INTELLECTUAL CAPITAL MANAGEMENT
7, 8 OCTOBER 2009
ZANJAN SCIENCE AND TECHNOLOGY PARK - INSTITUTE FOR ADVANCED STUDIES IN BASIC SCIENCES (IASBS)

Some applications of mathematical modeling methods in the Tajik language

Mizrob Ismoilov
Technological University of Tajikistan, Dushanbe, Tajikistan

Abstract: This article presents the applied mathematical modeling methods in some problems of Tajik language, as follows:

- Syllabification of arbitrary Tajik word;
- Morphological analysis of Tajik words formed on bases of all-embracing meaningful parts of the speech. Tajik words morphological synthesis of the given morphs set.
- Computer non-stress sounding of Tajik phrase. Automatic stressed syllables definition in the Tajik word; these objectives are the basis for the creation of automatic (computer) translators from the Tajik language into other languages and vice versa – from other languages in the Tajik language. Also these models can be used for creation of computer sounding of Tajik texts.

Keywords: Morphological analysis/synthesis, syllabification of words, non-stress sounding of Tajik phrase

INTERNATIONAL CONFERENCE ON
INTELLECTUAL CAPITAL MANAGEMENT
7, 8 OCTOBER 2009
ZANJAN SCIENCE AND TECHNOLOGY PARK - INSTITUTE FOR ADVANCED STUDIES IN BASIC SCIENCES (IASBS)

This article presents the applied mathematical modeling methods in some problems of Tajik language, as follows:

1. Syllabification of arbitrary Tajik word;
2. Morphological analysis of Tajik words formed on bases of all-embracing meaningful parts of the speech. Tajik words morphological synthesis of the given morphs set.
3. Computer non-stress sounding of Tajik phrase. Automatic stressed syllables definition in the Tajik word;

These objectives are the basis for the creation of automatic (computer) translators from the Tajik language into other languages and vice versa – from other languages in the Tajik language.

As it is known [1], words are constructed in two ways:

- The method of word composition by adherence of two or more root words;
- The method of word derivation, when to join a prefixes (to left) and postfixes (to right) to word stem.

Obviously, the vast majority of words are produced by using the second method.

There are also a large number of words created by applying both methods, as well as by combining root words with using infixes.

Syllabification of arbitrary Tajik word 1

Arbitrary Tajik word has the form

$$S = s_1 s_2 s_3 ... s_m , \qquad (1)$$

where, $s_i = (i = 1,2,3,...,m)$ letters of the alphabet can be represented [2] in the form of

$$S = S_1 \alpha_1 S_2 \alpha_2 S_3 \alpha_3 S_4 \qquad (2)$$

where

$$S_j = p_{1,j} p_{2,j} O_j q_{1,j} q_{2,j} ... q_{n,j} \qquad (3)$$

Here, $p_{i,j}\left(i = \overline{1,3}; j = \overline{1,4}\right)$ and $q_{i,j}\left(i = \overline{1,7}; j = \overline{1,4}\right)$ are possible prefixes and postfixes correspondingly, $\alpha_i\left(i = \overline{1,3}\right)$ - possible infixes, $O_j\left(j = \overline{1,4}\right)$ - the root word. Formula (2), and (3) are a model of arbitrary Tajik words.

Along with the morphological presentation of words in some problems there is need to present words in a sequence of syllables attached to each other. It concerns transfer of words to a new line, audio reading of texts, etc. Traditional Tajik words written with the the existing alphabet, based on the Russian Cyrillic alphabet and syllables are composed of species

Г, ГС, ГСС, СГ, СГС, СГСС (4)

where symbols Г and С means vowel and consonant to the letter respectively.

The algorithm, introduced in [3], allows to divide an arbitrary one-root word into correct syllables. Apart from the word itself this algorithm does not require any other information. However, this algorithm divides some multirooted words into the syllables incorrectly. Universal algorithm, which allows splitting any arbitrary word into syllables, requires a prior morphological analysis of a word, i.e. its presentation accordingly to formula (2) and (3).

Morphological analysis of Tajik words formed on bases of all-embracing meaningful parts of the speech. Tajik words morphological synthesis of the given morphs set 2

One of the characteristics of Tajik language is the fact that some groups of complex words are equal to a fragment of sentence in another language and translation according to the scheme

the word of input language - the word of output language

is not always possible. These words can be presented [1] as a fragment of Tajik sentence. Presentation of such words in the form of sentense fragments of English language was made in the papers [4, 5].

Another feature of the word-formation in Tajik language is that the regular morph joint to the word does not change the word stem (its former state), i.e. a normal concatenation of morphs occurs.

The object of morphological analysis of words is to provide the words (1) in morphologically disassembled form (2), (3) and attribution of grammatical categories (the definition of grammatical categories for all its morphs) to all its grammatical morphs.

The necessary attributes for the development of mathematical models of morphological analysis of words are

- the computer dictionary of word stems, indicating the parts of the speech, which they belong to.
- the data base of prefix, infix and postfix
- adherence rules of morphs to each other;

The word stem computer dictionary is based on an alphabetical order, as in the section of words beginning with the same letter, words are grouped by the number of letters.

In the Tajik language there are 13 simple prefixes (they begin from one of the letters *R { б, в, д, м, н, п, т, х }*, 10 complex prefixes, and 3 infixes (u, o, ma). Postfixes in the Tajik language, taking into account their omonimichnosti are more than 147.

The list of morphs includes two fictitious morphs "[" and "]" to denote the beginning and the end of words, respectively. In addition, a great number of words belonging to the particular part of the speech are taken as a generalized morph. For example, all nouns are combined into one morph.

The words related to the verb, depending on the category of time (past and present-future time) and at the presence or the absence of prefix *ме-* and *хаме-*in their composition are categorized into four separate generalized morphs.

For the word formation out of the word stems of specific parts of the speech the particular subset of the affix set is used. At the same time the certain number of morphs of the same subset can be affixed directly on

the right of every morph of this subset. Thus, the word formation in the Tajik language can be presented as a oriented graph (word formation tree).

The process of morphological words analysis is conveniently divided into several stages.

1) if the first letter of the word sample does not belong to the set *R б, в, д, м, н, п, т, х}*has the word does not contain a prefix, then - shift in the 5;
2) if the prefix is identified , it is cut from the word and then - shift to p.1);
3) go to the search for word stems in the computer dictionary;

As soon as the word stem has been found a generalized word stem $O_1 = PO$

(prefix + base).is composed. Then the part of speech [6], which the generalized word stem O_1 belongs to, is identified.

Analysis of postfixes can be carried out both by using separate models for the generalized word stem of specific parts of speech, and by means of the general model including all generalized word stems as separate morphs.

The overall model of morphological analysis postfix is a combination of rules of adherence morphologic type

$$K.\, m_k\left[\delta_k(m_1,1);\delta_k(m_2,2);...;\delta_k(m_N,N)\right]$$

Here, K is the number of rule, m_k is the morph with number k, $\boldsymbol{N}$ is a total number of morphs, $\delta_k(m_j,j)=1$ if morph m_j can directly join to morph m_k right and $\delta_k(m_j,j)=0$ in other case. The figure j in $\delta_k(m_j,j)$ means that if $\delta_k(m_j,j)=1$, then following the m_j morph should look for in rule number j.

Model (5) is a square NxN -th order matrix, which elements are zeros and ones.

Having identified postfixes of the analyzing words according to the model (5) and having assigned grammatical categories to all of their morphemes (prefix and postfix), we thereby accomplished morphological analysis of the studied words.

MORPHOLOGICAL SYNTHESIS is the inverse problem in relation to the morphological analysis and consists in the following:

a random suite of morphs is set

$$m_1^\circ, m_2^\circ, ...m_l^\circ \qquad (6)$$

with their grammatical characteristics.

It is required:

- to define the possibility (or impossibility) of synthesis of word from the given suite of morphs (6)
- if the synthesis is possible, then to implement it, otherwise to specify the reason of impossibility of the synthesis;

We must say that the number of morphs in the suite (6) can not exceed the number 13, because the word may contain no more than three prefixes, and not more than seven postfix. Symbols of the beginning "["and the end "]" of the words are also included here.

Comparing the elements of suite (6) with the elements of data base of prefixes, computer dictionary of word stems and elements of data base of postfixes we determine the existence and number of prefixes, word stems and postfixes, defined in a suite of morphs.

If the number of prefixes is more than three, or the word stem is missing, or the number of postfixes is more than seven, then the synthesis of words from the set (6) is obviously impossible.

Computer non-stress sounding of Tajik phrase. Automatic stressed syllables definition in the Tajik word 3

One of the important applications of computer modeling is audio reading texts. From our point of view, a dictionary with the elements

Syllable – Sound

is the most acceptable for Tajik language.

The feasibility of such approach is justified by the fact that

- the number of syllables in Tajik language is limited;
- the words are pronounced by syllables with little pauses between them;

Thus it is quite sufficient to have data base of syllables and their sound translation for the non-stress audio reading.

In [2,3] a simple algorithm for partitioning one-rooted Tajik words in syllables was proposed. Apart from the word itself this algorithm does not require any other information. Just type a certain amount of texts in the computer, make a partition of words into syllables, and thus create the database of syllables. For the first time this database for the Tajik language was executed in [7]. Sound of syllables at the current capabilities of computer technology does not require any special work.

In spite of being very simple, this algorithm is not universal yet: there are a great number of multi-syllable words with the consonant letter at the end of one root and another root, which is attached directly to right of it, begins with the vowel letter. Such words this algorithm divides into syllables incorrectly.

Regarding to this, the universal algorithm was proposed. Before applying this algorithm it is necessary to perform morphological analysis of multi-rooted words and to present the word in the form of formulas (2), (3) and to split each word type (3) into syllables independently.

[1] Sergey Arzumanov, Orif Jalolov (1969). “Tajik Language”. Dushanbe: Irfon.

[2] Mizrob Ismoilov. (1994). “Fundamentals of automatic morphological analysis of Tajik words”. Dushanbe: NPICenter.

[3] Mizrob Ismoilov. (2000). “Algorithm of automatic syllabification of Tajik words”. Dushanbe: Academy of Sciences of Tajikistan, Vol.43, No 3, pp 95-99.

[4] Mizrob Ismoilov, Shoira Pulatova. (2007). “Interlanguage and Intralanguage normalization groups of Tajik words, formed on basis of all-embracing meaningful parts of the speech. Dushanbe: NPICenter, No. 5 (1762).

[5] Shoira Pulatova. (2008). “Interlanguage and Intralanguage normalization groups of Tajik words formed on bases of the verb”. Dushanbe: NPICenter, No. 05 (1773).

[6] Mizrob Ismoilov, Shoira Pulatova, Yusuf Nabotov. (2008). “Database of affixes of Tajik language”. Dushanbe: NPICenter, No. 08 (1776).

[7] Zafar Usmanov, Abdugani Abduhamidov, Mizrob Ismailov. (2008). “Statistical regularity of syllabic composition of Tajik language ”. Dushanbe: Academy of Sciences of Tajikistan, Vol.45, No.5-6.