SYNCHRONIZED LINEAR TREE FOR MORPHOLOGICAL ANALYSIS AND GENERATION OF THE KAZAKH LANGUAGE

A. Sharipbay *G.Bekmanova*
sharalt@mail.ru *gulmira-r@yandex.kz*

B.Yergesh *A.Mukanova*
b.yergesh@gmail.com *asel_ms@bk.ru*

L.N Gumilyov Eurasian National University

ABSTRACT

The paper describes representation of the problem of inflection / derivation. To solve these problems can be used synchronized linear tree. In this case it will act as an a switching circuit switching excitement subavtomat transforming from one state to another.

1 Introduction

Agglutinative languages (lat. Agglutinatio — combine, stick) are languages that have a system in which the dominant type of inflection is the agglutination ("sticking") of different formants; these can be either a prefix or a suffix and have only one meaning[1].

The Kazakh language is part of the Turkic group of languages; this language group can be classified as an agglutinative language. Words in the Kazakh language contain many word inflections; inflections are formed by adding suffixes and endings to words. Suffixes and endings are attached in a strict sequence and words in the Kazakh language vary in number, case, and person. A possessive form in Kazakh exists as it does in the English language

The authors used various methods for morphological analysis and generation of the words in the Kazakh language. Here we show how to use a synchronized linear tree for solving the problems of inflection/derivation and morphological analysis.

2 Synchronized linear tree

Synchronized linear tree is composed of sub-layers of neurons. Each synchronized sub-layer corresponds to the wave front processing. Neurons of the first sub-layer correspond to the first letter of the word, the second - the second and so on. The total number of sub-layers is the maximum number of letters in a word. [2]

The internal structure of the lemma in a synchronized linear tree is as follows (Table 1).

Table 1 Notation in the internal structure of the lemma

Symbol	Notation
#	the space between words
<	word beginning
>	word end
!	morphological information for the inflection

For example, we construct a synchronized linear tree for the word “oqushy – pupil” (lemma). Receptor is excited at the beginning

of the word at symbol "<". Further it changes to “o”. When giving the symbol “o”, further sequentially “<oq”, “<oqu”, “<oqus”, “<oqush”, “<oqushy”, “<oqushy>”(Fig. 1).
When giving about symbol, further it is consecutive

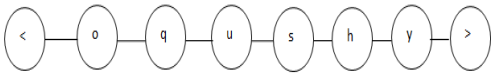

Figure 1. Synchronized linear tree for lemma “oqushy”.

We will consider features of lemmas. The symbol “<” denote the first special symbol of the lemma. The word form beginning and its feature are indicated by different special symbols to decrease the size of the search tree that can increase the operating speed of the sequential computing system. However, to solve the problem of inflection on a parallel computing system it would be enough to limit oneself to distinction of the special symbols “!” and “>”.
Figure 2 shows an example of the structure of relations of the lemma, which defines the following features: noun – “*ze!” Animate – “*ja!”, the last vowel symbol “y” [3].

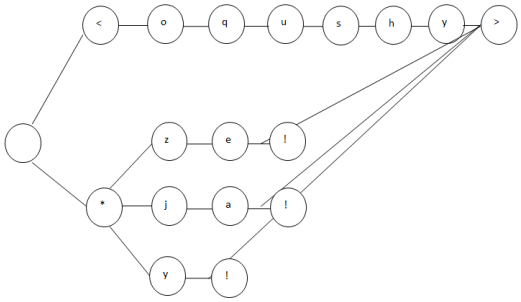

Figure 2. Synchronized linear tree for the lemma and its morphological information

In this tree, depending on the semantic features the trajectory of inflection is chosen First of all has value the sign of a parts of speech, according to it, we inflect the word according to rules. Next animate - a sign of inherent to noun. The fact that the personal endings are added only to animate nouns, i,e. you can say *“Men oqushymyn - I am a pupil”*, but we cannot say *“Men kurekpin – I am a shovel”*. Each part of speech has semantic features, i.e. features which are allocated at words on the basis of their meaning. Therefore, they have to store in a part of the morphological information.
In that case it is possible to present a lemma and its morphological information in the form of synchronized linear tree.
The switching of the sub-automat’s states will take place with issue of special commands to the synchronized linear tree input. These commands will be recognized by the synchronized linear tree and transformed to the gradient value at the output of neurons-effectors corresponding to them. This will cause excitation or deceleration of neurons corresponding to the lemma’s states.
In the case of excitation of these neurons further by the rules of inflection will be made the formation of new word forms. The last sound “y” specifies the type of the end for a noun.
At addition more than one termination it is analyzed, than the word basis terminates, the sign on which the first termination is added is allocated, the added termination is analyzed further, the sign on which the following termination etc. is added is allocated. In case of lack of one of the terminations (the zero termination) the subsequent termination is added to that precedes it.
When adding more than one ending it is analyzed the word basis end, then allocated feature by which the first end. The added ending is analyzed further, the sign on which the following termination etc. is added is allocated. Then allocated the feature by which the following is added the next end, and so on. In case of lack of one of the endings (zero

ending) the subsequent ending is added to that precedes it.

Thus, from all the possible combinations of word endings using neural network is automatically filled the dictionary of word forms.

Similarly, we can present a model of morphological analysis. The only difference is that in the first case, we store the initial dictionary of forms of words, and the second dictionary of word forms.

It is similarly possible to present model of the morphological analysis. The difference is, in the first case we store the dictionary of initial forms of words, and in the second the dictionary of word forms.

The internal structure of the word form in a synchronized linear tree is as follows (Table 2).

Table 2 Notation in the internal structure of the word form

Symbol	Notation
#	the space between words
<	word beginning
>	word end
$	word form feature (case, number, conjugation etc.)

We will consider the example for the word "bala - a child" (lemma) and two its forms "balam - my child", "balan - your child" (in the Kazakh language animate nouns change by persons using personal endings). The receptor is excited at the symbol beginning of the word "!". Further we pass into a state of "b." when submitting the symbol "b", further consistently "<ba", "<bal", "<bala", and then simultaneously two substates "<balam |» and «<balan |» (Figure 3).

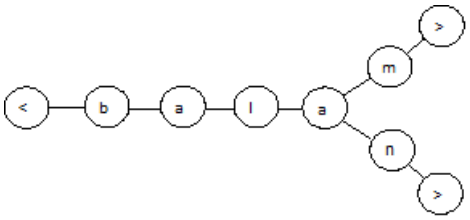

Figure3 . Synchronized linear tree for word forms

We will consider the features of word forms. Symbol "*" denote the first special symbol feature word forms, protruding the same role as the symbol "<" for the word forms. Beginning of word form and its feature denote different special symbols for reducing the size of the search tree, it is may increase the speed of operation of a consistent computer system. However, for solving the problem of inflection on parallel computing system, it would be suffices to restrict the distinction of special symbols "!" and ">". Figure 4 shows an example of the structure of relations of the lemma, which defines the following features: noun–«**ze*!» , personal ending, first person, possessive ending, second person. When applying to the lemma of the word "<balam>" it passes to the excited substate "<balam>", «**ze*!», «**j1bje*!», and when applying to the word "balan" in the excited substate "<balan>"«**ze*!», «**j2bje!*».

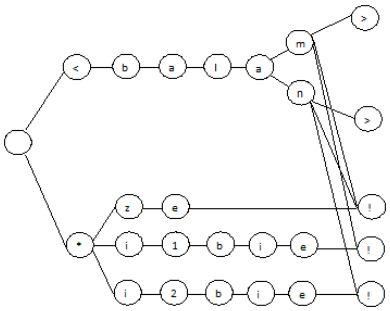

Figure 4 . Synchronized linear tree for word forms and their morphological information.

3 The methods of morphological analysis

The method of morphological analysis, when the disc is stored whole dictionary is the declarative. In this method of the morphological analysis in the dictionary stores all the possible word forms of each word and the corresponding morphological information. In this case the problem of the morphological analysis is to search the word form in the dictionary and rewrite from the dictionary morphological information. Since the amount of different word forms of each word is quite large, declarative method requires a considerable amount of computer memory system, which is accompanied by some difficulties related to the creation and support of the dictionary, as well as redundant information. The advantages of this method include high-speed analysis and universality in relation to the set of all possible word forms[4].

Procedural method presupposes preliminary systematization of morphological knowledge about the natural language and the development of algorithms for assignment morphological information separate word form. Procedural morphological analyzer consists of the following steps: selection bases of in the current word form, its identification, the attribution word form of the corresponding list of morphological information[4]. The disadvantages of of this method are the high complexity of compiling dictionaries compatibility, that is intractable and not fully automatable task for languages, which is characteristic of a large amount of stopwords. The implementation of this method takes much up less memory space, but this increases time of morphological analysis word forms by dividing into components and application procedures compatibility.

When using a procedural method of morphological analysis algorithm becomes much more complicated. The fact is that, for example, for nouns personal ending (first person, singular) the "m" is included in the other endings "-min" "-miz" and others.

If the declarative method of analysis has not given desired results is used a procedural method.

We will consider the example for the word "bala - a child" (lemma) and two its word forms "balamnyn - my child", "balamyn - I am a child". In the first case to the base added two endings: "-m" is possessive ending (first person, singular) , and "-nyn" is case ending. In the second case to the base added one ending "-myn" is personal ending (first person, singular). The search algorithm should include any possible amount of attached endings and accumulation of morphological information.

Below describes the algorithm of the search word and its morphological information:

1. Word is read;
2 Opens dictionary initial forms and it searches for the read words;
3 If the word is found, then go to step 8, otherwise step 4;
4. Word in the cycle is read symbol by symbol, starting with the last symbol that it turns out looking in the ending's dictionary;
5. If the ending found, the rest are looking for in the dictionary of the initial forms;
6. Save morphological information of the words;
7. If this word is not found, go to step 4, otherwise go to step 8;
8. End.

The combined method. In systems the real degree of difficulty is more often used combined version of the morphological analysis. Here we use a dictionary of word forms and dictionary basics. On the first stage carried a dictionary lookup word forms, as in a declarative method, and in the case of a successful search in this analysis is completed. Otherwise activated dictionary bases and procedural analysis method.

4 Conclusion

In this paper we have presented synchronized linear tree to solve the problems of morphological analysis and generation of words in the Kazakh language. Also discussed some methods of morphological analysis and presented an algorithm for morphological analysis.

Automatic word formation and inflection can be used in the systems of automatic speech generation or recognition, as well as in rather traditional area of Kazakh language study, in orthographic correctors, translators, morphological analysis, as one of integral components of the given process is the training of reading skills, i.e. the reading of a written text.

The formed dictionaries can be issued as orthographic dictionaries. The obtained formalizations, methods and algorithms can be used in NLP systems (orthographic correctors, translators, training systems), Kazakh speech recognition and synthesis systems, as well as in semantic search systems.

3 References

[1] **BODMER, FREDERİCK. ED. BY LANCELOT HOGBEN,** The Loom of Language. New York, W.W. Norton and Co., 1944, renewed 1972, pages 53, 190ff. ISBN 0-393-30034-X

[2] ***D.E.SHUKLİN.*** Structure of semantic neural net parsing text meaning in real time // Cybernetics and System Analysis. Kiev. Glushkov Institute of Cybernetics, 2001. - *N*o 2. p. 43-48.

[3] ***SHARİPBAEV A., BEKMANOVA G.*** The synthesis of word forms of Turkic language using semantic neural networks // Abstracts «Modern problems of applied mathematics and information technologies – Al Khorezmy 2009». – Tashkent, 2009. – P. 145.

[4] ***L.V. NAYHANOVA, I.S. EVDOKİMOVA.*** Methods and algorithms for translation of natural language queries to a database in SQL-queries. - Ulan-Ude: Publisher ESSTU, 2004 – p. 148. In Russian.