UDC 004

AUTOMATIC PROCESSING OF TEXT IN NATURAL LANGUAGE

*©**Kochkonbaeva B.**, Osh Technological University named by M.M. Adyshev,*
Osh, Kyrgyzstan, buajar@mail.ru
*©**Aldosova A.**, Osh Technological University named by M.M. Adyshev,*
Osh, Kyrgyzstan, a_aldosova81@mail.ru

АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

*©**Кочконбаева Б. О.**, Ошский технологический университет*
им. акад. М. М. Адышева, г. Ош, Кыргызстан, buajar@mail.ru
*©**Алдосова А. Ю.**, Ошский технологический университет*
им. акад. М. М. Адышева, г. Ош, Кыргызстан, a_aldosova81@mail.ru

Abstract. In this article, questions of artificial intelligence, in particular, automatic processing in natural language texts are considered.

As well as types of wordform analysis are considered and an algorithm for finding the initial form of the word is proposed.

Аннотация. Рассматриваются вопросы искусственного интеллекта, в частности, автоматическая обработка текстов на естественном языке.

Рассматриваются типы анализа текстовой информации и предлагается алгоритм поиска исходной формы слова.

Keywords: computer linguistics, natural language, dictionary, morphology, Kyrgyz language, algorithm.

Ключевые слова: компьютерная лингвистика, естественный язык, словарь, морфология, кыргызский язык, алгоритм.

Introduction

Following the advent of computer technology, problems of text processing arose. Information technology and research in the field of artificial intelligence are evolving every day, but there is as yet no satisfactory solution to most problems of processing the text of a natural language. Computer linguistics is a branch of science that studies the application of mathematical models to describe linguistic regularities. It can be divided into two large parts. One of them studies the methods of applying computer technology in linguistic studies - the application of known mathematical methods (for example, statistical processing) to identify patterns. The discovered regularities are used by another part studying the issues of comprehending texts written in natural language - the creation of mathematical models for solving linguistic problems and the development of programs that operate on the basis of these models. This part of computer linguistics is closely related to the section on artificial intelligence, which is developing text processing systems in natural language.

The general scheme of text processing (Figure 1) is invariant with respect to the choice in natural language. Regardless of the language in which the source code is written, its analysis passes through the same stages. The first two stages (splitting the text into separate sentences and into words) are practically the same for most natural languages. The only thing that can affect the specific features of the chosen language is the processing of word abbreviations and the processing of punctuation marks (more precisely, determining which of the punctuation marks are the end of the sentence and which are not).

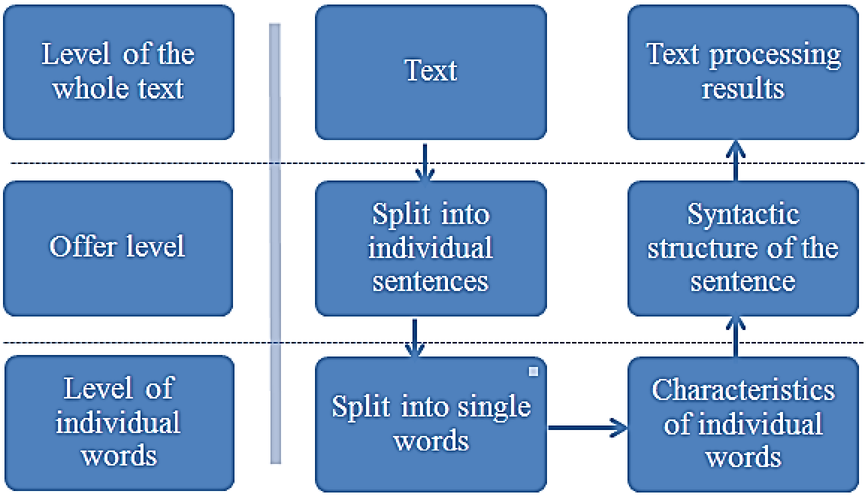
Level of the whole text
Text
Text processing results
Offer level
Split into individual sentences
Syntactic structure of the sentence
Level of individual words
Split into single words
Characteristics of individual words

Figure 1. General scheme of text processing.

The next two stages (characterization of individual words and syntactic analysis), on the contrary, depend heavily on the chosen natural language. The last stage (semantic analysis) also depends little on the chosen language, but this is manifested only in general approaches to analysis.

Substantial support in carrying out linguistic research is provided by programs that automatically find the required word forms in the texts under study. For this, special programs should be compiled that perform an automatic search for word combinations.

An important part in the automatic processing of texts in natural language is the technology of finding the basis of a word, an algorithm similar to it for purposes that allows one to determine that some chain of word forms constitute one inflectional group. A program capable of performing these operations includes the morphological analysis of the word in automatic mode.

The problem of processing texts in the Kyrgyz language, "understanding" of the language by the computer, is an actual task at the given time. Among the many tasks that are reduced to solving this problem, you can name such as communication with a computer in a natural language (question-answer systems), information search, machine translation, extracting useful information from texts, etc.

It is enough routine work - to analyze the style of any author for his work. With the help of the automatic word decomposition into morphemes and statistical data, it becomes possible to automatically analyze author texts and compose ready-made concordances.

For this purpose, a study was made of the morphology of the Kyrgyz language. A correct understanding of the composition of the word, the ability to determine its constituent components is of great importance in the study of language. The word reflects the features of the language structure, its lexical-semantic and functional-grammatical laws.

The Kyrgyz language is characterized by relative regularity, positional and grammatical stability of the morphological structure of various word forms. The formation of words is the successive adherence to the basis of the word grammatical particles - affixes (for example, кыргыз + дар + сыңар + бы).

Analysis of individual words

This stage of processing includes morphological and morphemic analyzes of words. The input parameter is the text representation of the source word. The goal and result of the morphological analysis is the definition of the morphological characteristics of the word and its basic word form. The list of all the morphological characteristics of words and the permissible values of each of them depends on the natural language. Nevertheless, a number of characteristics (for example, the name of a part of speech) are present in many languages. The results of the morphological analysis of the word are ambiguous, which can be traced to a lot of examples.

There are three main approaches to conducting the morphological analysis. The first approach is often called "clear" morphology; The second approach is based on a certain system of rules, based on a given word defining its morphological characteristics; in contrast to the first approach, it is called "fuzzy" morphology [2]. The third, probabilistic approach is based on the compatibility of words with specific morphological characteristics; It is widely used in the processing of languages with the strictly fixed order of words in the sentence and is practically not applicable when processing texts in inflectional languages. Let's consider all three methods of morphological analysis in more detail.

The dictionary of the basics, which we collected contains the main word forms of the words of the Kyrgyz language. There is a system of rules with which you can build all forms of a given word, starting from the initial word form and the code corresponding to it. In addition to constructing each word form, the system of rules automatically puts in correspondence with its morphological characteristics. When carrying out a clear morphological analysis it is necessary to have a dictionary of all words and all word forms of the language. This dictionary at the input takes the form of a word, and at the output gives out its morphological characteristics. This dictionary can be built on the basis of the dictionary of the Kyrgyz language by an obvious algorithm: to sort through all the words from the dictionary, for each of them to determine all possible word forms and to put them into the emerging dictionary.

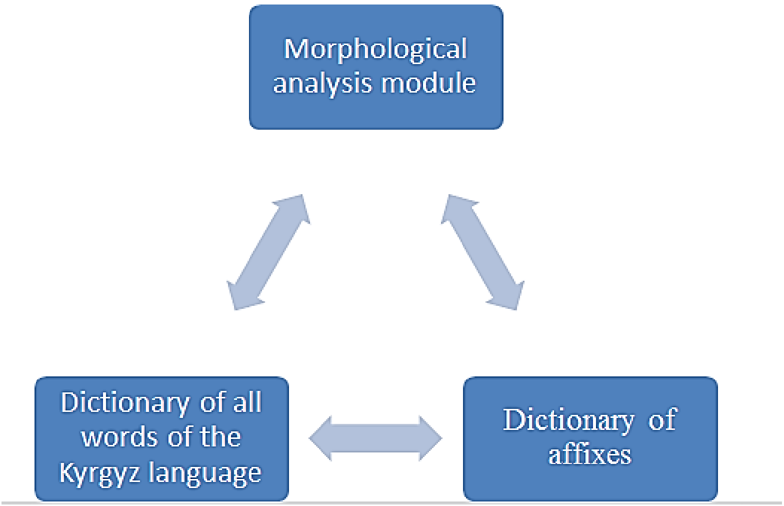
Morphological analysis module
Dictionary of all words of the Kyrgyz language
Dictionary of affixes

Figure 2. Morphological analysis based on the dictionary

With this approach, to perform a morphological analysis of a given word (Figure 2), it is simply necessary to find it in the dictionary, where the exact, "finally known" values of all its morphological characteristics are already stored. For the same input word, several variants of the values of its morphological characteristics can occur at once.

Unfortunately, this method is not always applicable: words entering the input may not be included in the dictionary of all word forms. Such a situation can arise due to errors in the input of the source text, due to the presence of names in the text, etc. In the case when the method does not give the desired result, fuzzy morphology is applied.

The purpose of the morphemic analysis of the word is to divide the word into roots and endings. In the dictionary of morphemes of the Russian language the division of each word into separate parts is indicated, but the types of each of them are not specified - which of them is a prefix, what is the root, etc. The set of all roots of the words of the Russian language is open, but the set of all possible prefixes, suffixes and endings is limited; In addition, it is known that in any word first go prefixes, then roots, then suffixes and endings. Therefore, based on the dictionary morpheme of the Russian language, you can build another dictionary that will contain not only the breakdown of each word into parts, but also the type of each of them. In this case, for carrying out the morphemic analysis of the word, it is necessary to refer to this dictionary.

The morphemic analysis is not limited to references to the dictionary. In a situation where the word is not in the dictionary, it is possible to conduct a direct analysis based on the standard structure of the words of the Russian language (prefix-root-suffix-ending) and the set of all consoles, suffixes and endings.

Let us return to the morphological analysis of the word in the situation when it was not possible to determine the characteristics of the word with the help of methods of clear morphology, but it was possible to break it apart. The presence of certain lexemes can determine the morphological characteristics of the word: you can build a system of rules that will rely on the presence or absence of any parts and give out one or more assumptions about the morphological parameters. Such a set of rules can be constructed in two ways. The first is based on the morphemic analysis of words contained in the dictionary of all word forms, and their morphological characteristics. We consider this problem more formally: pairs of values are known, consisting of the morphemic structure of the word and its morphological characteristics. This is nothing more than the "input" and "exit" of the rules system, which, by the morphemic structure of the word, will determine its morphological characteristics. The task of constructing such a system of rules can be solved with the help of a self-learning system (Figure 3). For its implementation, decision trees, programming based on inductive logic (ILP, Inductive Logic Programming) or other algorithms can be used.

The second approach is to create a set of rules manually. By and large, its implementation is nothing more than writing an expert system of diagnosing type.

The probabilistic method [3] of morphological analysis of words is as follows. The same word form can belong to several grammatical classes at once. For each word form, all its grammatical classes are defined, as well as the probability of its relation to each of these classes. This is done on the basis of some set of documents, where each word is preceded by a grammatical class. After that, the probabilities of combinations of certain grammatical classes for words standing side by side - for twos, triples, quads, etc. — are calculated. On the basis of these numbers, words can be analyzed, but for him, it is necessary not only the word itself but also the words next to it.

Two important observations need to be made. First, the probabilistic method is applicable only for languages that have a clearly fixed word order in the sentence. If the order of words can be changed, then all possible combinations of grammatical classes will be almost equally probable. Secondly, if the first two methods of analysis (clear and fuzzy morphology) accept individual words at the input, then the probabilistic method, on the contrary, accepts either the entire sentence at the input or at least several words that stand side by side.

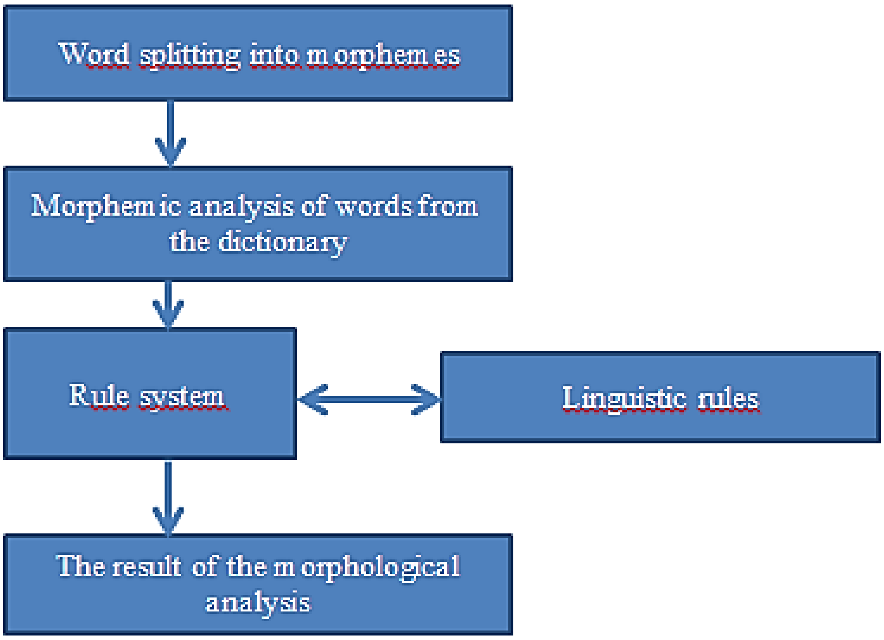
Word splitting into morphemes
Morphemic analysis of words from the dictionary
Rule system
Linguistic rules
The result of the morphological analysis

Figure 3. Fuzzy morphological analysis

Morphological analysis algorithm

The normalization module in the process during its work performs the following sequence of steps:

1-Step: A word is searched in the dictionary of the initial forms. If a word is found in the dictionary, go to step 5.

2-step: The word is read symbol-by-symbol in the reverse order (beginning with the end of the word). If the word is finished, then the algorithm's work ends. Based on the current list of affixes, a list of hypothetical affixes is formed.

3-step: All hypothetical affixes in the affix dictionary are searched. All found affixes are added to the list of affixes. If no new affix is found, go to step 2.

4-step: The initial part of the word is searched in the dictionary of the initial forms. If no word is found, go to step 2.

5-step: The result is added found foundations and a concomitant set of affixes. Go to step 2.

After normalization, for each word found, its morphological characteristics are calculated on the basis of its affixes and the morphological class of the stem.

Conclusion

Recently, thanks to the development of document management systems, the availability of a set of constantly updated legal guides, and a number of other factors, there is an accumulation of arrays of specialized (but not formalized) text documents. By analogy with structured information, when the development of analysis tools has resulted in the emergence of data warehouses, the development of document management systems over time may require the creation of full-text storage facilities that enable comprehensive analysis and research of non-formalized texts in natural language.

References:

1. Sadykov, T., Zhumalieva, G. E., Tumonbaev, M. Zh., & Sharshembiev, B. (2015). Computer linguistic bases of the Kyrgyz language. *Bishkek.*

2. Manning, C. D., Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

3. Gryaznukhin, T. A., Darchuk, N. P., Kritskaya, V. I., & Malovitsa, N. P. (1999). Syntactic analysis of scientific text on a computer. Kiev: *Naukova Dumka,* 272.

4. Hunt, E. (1978). Artificial Intelligence: Trans. with English. A. A. Belova, Yu. I Kryukova, ed. V. L. Stefanyuka.b: *World.* 558.

Список литературы:

1. Садыков Т., Жумалиева Г. Е., Тумонбаев М. Ж., Шаршембиев Б. Компьютерные лингвистические основы кыргызского языка. Бишкек, 2015.

2. Manning C. D., Manning C. D., Schütze H. Foundations of statistical natural language processing. MIT press, 1999.

3. Грязнухин Т. А., Дарчук Н. П., Крицкая В. И., Маловица Н. П. и др. Синтаксический анализ научного текста на ЭВМ. Киев: Наукова думка, 1999. 272 с.

4. Хант Э. Искусственный интеллект: Пер. с англ. Д А Белова, Ю И Крюкова, под ред. ВЛ Стефанюка.Ь.: Мир. 1978. 558 с.

Работа поступила
в редакцию 01.06.2018 г.

Принята к публикации
04.06.2018 г.

Cite as (APA):

Kochkonbaeva, B., & Aldosova, A. (2018). Automatic processing of text in natural language. *Bulletin of Science and Practice, 4*(7), 216-221.

Ссылка для цитирования:

Kochkonbaeva B., Aldosova A. Automatic processing of text in natural language // Бюллетень науки и практики. 2018. Т. 4. №7. С. 216-221. Режим доступа: http://www.bulletennauki.com/kochkonbaeva (дата обращения 15.07.2018).