разработка предлагает те же основные функции, что и системы-аналоги, но отличается тем, что подсчёт баллов будет вестись и по правилам ACM ICPC для студенческих олимпиад по программированию, и по правилам IOI для школьных олимпиад по информатике. Разработка этой системы выводит проведение Республиканской олимпиады школьников по информатике в Кыргызстане на уровень международных олимпиад.

Список литературы

1. Сайт проведения международных студенческих командных соревнований: https://icpc.baylor.edu
2. Сайт проверочной системы КРСУ http://olymp.krsu.edu.kg/
3. Сайт проверочной системы ТИМУС http://acm.timus.ru/
4. Корнеев Г.А., Станкевич А.С. Методы тестирования решений задач на соревнованиях по программированию. //Труды II межвузовской конференции молодых ученых. – СПб.: СПбГУ ИТМО, 2004. С.36-40.
5. Станкевич А.С. Общий подход к подведению итогов соревнований по программированию при использовании различных систем оценки. //Компьютерные инструменты в образовании. №2, 2011. С 27-38.
6. Макиева З.Дж., Каримова Г.Т., Макаева А.Д. Требования к веб-ориентированной информационной системе по секторальной квалификационной рамке Кыргызской Республики для ИКТ направлений. //Теоретический и научно-методический журнал «Вестник университета (ГУУ)» ФГБОУ ВПО «Государственный университет управления», Москва, № 19/2014. С.167-172.

UDC 025.4.03

APPLICATION OF MORPHOLOGICAL MARKUP OF KAZAKH LANGUAGE TO AUTOMATED FILLING OF THE ONTOLOGY OF FACTOGRAPHIC RETRIEVAL SYSTEM

Madina Mansurova, *al-Farabi Kazakh National University, Almaty, Kazakhstan,.*
Kairat Koibagarov, *Institute of Information and Computational Technologies, Almaty, Kazakhstan.*
Vladimir Barakhnin, *Institute of Computational Technologies SB RAS, Novosibirsk, Russia, Novosibirsk State University, Novosibirsk, Russia.*
Madina Soltangeldinova, *al-Farabi Kazakh National University, Almaty, Kazakhstan.*
Serzhan Berdibekov, *al-Farabi Kazakh National University, Almaty, Kazakhstan.*

Abstract. This work is concerned with the development of the parser for automation of morphological markup of the texts of the Kazakh National corpus. The parser includes the lexical and morphological analyzers to perform the morphological markup of the texts. The task of a lexical analyzer is to determine the boundaries of sentences, to display words, identifiers and punctuation marks. The morphological analyzer performs the search for words in the dictionary (which is a separate database) and determines their morphological parameters. At the output of the morphological analyzer, we have a list of lemmas (a normal form of the word), affixes and morphological characteristics of the word. Morphological markup of the texts is a stage of automatic text processing, which allows to use the marked texts to solve the different problems of Natural Language processing. This paper describes the application of morphological parser of the Kazakh language to automated filling of the ontology of factographic retrieval system.

Keywords: Morphological Parser, Morphological Markup, Factogaphic Retrieval, Facts Extraction, Automated Ontology Filling.

ПРИМЕНЕНИЕ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА КАЗАХСКОГО ЯЗЫКА ДЛЯ АВТОМАТИЗИРОВАННОГО НАПОЛНЕНИЯ ОНТОЛОГИИ ФАКТОГРАФИЧЕСКОЙ ПОИСКОВОЙ СИСТЕМЫ

Мансурова М.Е, *Казахский национальный университет имени ал-Фараби.*
Койбагаров К.Ч, *Институт информационных и вычислительных технологий МОН РК.*
Барахнин В.Б., *Институт вычислительных технологий СО РАН.*
Солтангельдинова М., *Казахский национальный университет имени ал-Фараби.*
Бердибеков С., *Казахский национальный университет имени ал-Фараби.*

Аннотация. Данная работа посвящена разработке анализатора для автоматизации морфологической разметки текстов корпуса казахского языка. Для осуществления морфологической разметки используются лексический и морфологический анализаторы. Задачей лексического анализатора является определение границ предложений, выделение слов, идентификаторов и пунктуационных маркеров. Морфологический анализатор выполняет поиск слов в словаре казахского языка и определяет их морфологические параметры. На выходе морфологического анализатора мы получим список лемм (нормальная форма слова), аффиксов и морфологических характеристик слова. Осуществляемая с помощью разработанного анализатора морфологическая разметка является этапом автоматической обработки текста, которая позволяет осуществлять поиск нужных пользователю слов, форм слова, лексических конструкций и т.д. В данной работе описывается применение модуля морфологического анализатора для автоматизированного наполнения онтологии фактографической поисковой системы.

Ключевые слова: морфологический анализатор, морфологическая разметка, фактографический поиск, извлечение фактов, автоматизированное наполнение онтологии.

Introduction

In Turkic philology, there are quite many investigations on this problem for cognate languages based on different conceptual approaches [1; 2; 3; 4; 5]. Analysis of the study on open publications in the field of the technology of morphological analysis of the Kazakh language word forms shows that in this field there are practically few investigations.

In the period from 1970 to 2000, the publications in the field of the Kazakh language morphology were largely theoretical. Since 2006 there were the valuable publications in international journals: Jonathan North Washington (2006) "A Novel Approach to Delineating Kazakh's Five Present Tenses: Lexical Aspect"; G. Altenbek and Wang Xiao-long (2010) "Kazakh Segmentation System of Inflectional Affixes"[6]; Zafer H.R., Tilki B., Kurt A., Kara M. “Two-level description of Kazakh morphology” (2011) [7]. Particular attention should be given to the work of Sharipbaev A.A. "Intelligent morphological analyzer based on semantic networks" (2012) [8]. All the listed works deal with some aspects in the field of morphology and syntax of the Kazakh language and have a theoretical character of investigations. In this relation, creation of a module for morphological analysis of the Kazakh words at a high processing rate is actual.

The remaining part of the article is structured as follows: Section 2 describes of development of the parser for automation of morphological markup of the Kazakh-language texts. Section 3 provides a detailed description of the technology of automated factographic retrieval system ontology filling. This technology contains extracting keywords from corpus of texts with similar topic for following using these keywords as possible values of entity’s attributes. Next, Section 4 describes the practical results. Finally, the article is completed by a relevant discussion and further research directions.

Development of the parser for automation of morphological markup of the texts of the Kazakh National corpus

Peculiarities of the Kazakh morphology

The Kazakh language refers to the class of agglutinative languages and together with Uzbek, Kyrgyz, Bashkir, Tatar, Azerbaijani, Turkish and other languages forms a Turkic linguistic family. Agglutinative languages are characterized by a consecutive addition of suffixes or ending bearing a grammatical meaning to an unchangeable root or stem having a lexical meaning.

The sequence order of affixes is strict. For example, for nouns first a suffix is added to the stem and then the ending of the plural followed by a possessive ending, then comes a case ending and finally the ending of the conjugation form (which is only added to animate nouns) [9; 10].

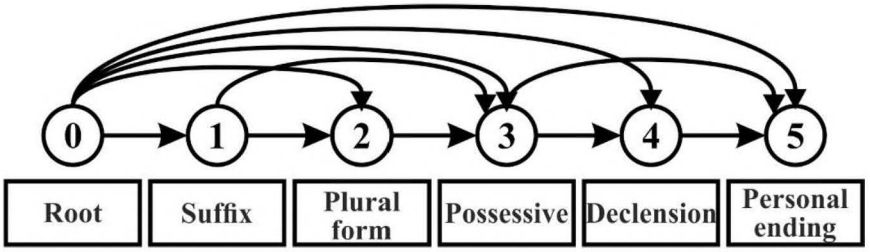

Fig. 1. Rules of affixes attachment for nouns

In the Kazakh language, there exists the law of vowel harmony: harmony between vowels and consonants of affix and the sounds of root. Vowels harmonize according to the hardness – softness principle and, as far as consonants are concerned, there is harmony between the final sound of root and the first sound of affix.

Apart from the three main rules of vowel harmony, it is necessary to take into account the following rules of exclusion:

1. The rule of removing a voiceless consonant in the affix being added if there are two voiceless consonants in the word ending. For example, *журналист + мер* (*zhurnalist + ter*) → *журналистер* (*zhurnalister*), *экстремист+мер* (*ekstremist + ter*) → *экстремистер* (*ekstremister*).

2. The law of vowel harmony is not observed in the following cases of the following affixes:

a) for affixes *мен, пен, бен* (*men, pen, ben*): *қаламмен* (*qalammen*); *нікі, дікі, тікі* (*niki, diki, tiki*): *баланікі* (*balaniki*);

b) for loan words with ending: : *рк, нк, кс, кт* (*rk, nk, ks, kt*): – *пункте* (*punkte*).

The law of omitting vowels “і”, “ы” (“i”, “y”) in the root, when adding a possessive affix “і”, “ы” (“i”, “y”). For example: *әріп – әрпі* (*arip – arpi*), *қойын – қойны* (*koiyn – koiny*).

The structure of a linguistic parser

Figure 2 presents the developed by us linguistic parser consisting of four analyzers (lexical, morphological, syntactic and semantic). The analyzers are successively arranged one after other, the output flow of one analyzer serving as an input for the following one.

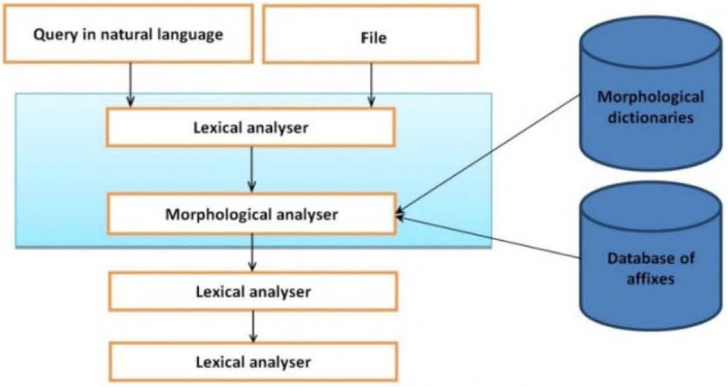

Fig. 2. The structure of linguistic parser

The task of a lexical analyzer is to determine the boundaries of sentences, to display words, identifiers and punctuation marks. The morphological analyzer performs the search for words in the dictionary (which is a separate database) and determines their morphological parameters (e.g., part of speech, number, case, etc.). A syntactic analyzer constructs a syntactic graph of a sentence. In this work, we study two analyzers – a lexical and morphological analyzer.

A lexical analyzer is a program of initial analysis of a natural text presented in the form of a chain of Unicode symbols. The output information is needed to be processed further by morphological and syntactic analyzers.

The tasks of a lexical analyzer include:

Division of the input text into words, numbers, disjunctives, etc.

Isolation of idioms not having word changing variants;

Isolation of proper names, FNP (family name, name, patronymic) when the name and patronymic are presented by initials;

Isolation of e-mail addresses and file names;

Isolation of sentences from the input text.

The procedure of isolation of words, numbers and punctuation marks is quite evident. After reading – out the next paragraph in turn, the graphematic analyzer parses tokens and assigns the corresponding graphematic characteristics to them. At this stage, tokens are isolated according to blanks and punctuation marks. However, of the greatest difficultly is determination of the beginning and the end of sentence. A lexical analyzer contains a heuristic mechanism for determination of the sentence boundaries and the result of lexical analysis is not only an array of lexemes, but also indicators of the beginning and end of the current sentence in the text.

It is not simple to find the end of a sentence either, as it may seem. An exclamation or question mark is sure to indicate the end of a sentence, but a point may be put after an abbreviation and in the middle of a decimal fraction. One must also take into account complex units of measurement (кв.м, км/час), internet – addresses (http://yandex.ru), ordinal numbers written out in figures (1917- ж.), markup names (Kassymov K.S). Let us consider the following fragment:

1917 ж. 21-26 шілдеде Покровский С.И. Орынборда болған «Бүкілқазақтық» съезде (1917 *zh.* 21 – 26 *shildede Pokrovckiy S.I. Orynborda bolghan "Bukilqazaqtyq" syezde*).

There are four points and only the fourth points indicates the end of the sentence. Therefore, special test units are introduced into the analyzer. In particular, the simplest the simplest check: a lexeme just before the point must contain at least one vowel. This test makes it possible to choose many abbreviations used with a point in the end: *ж., т.е.к., тг., ф.-м.г.д., қ.* (*zh., t.g.k., f.-m.g.d., q.*).

Of course, suck a test does not guarantee the correctness of the result. On the one hand, an abbreviation or a number with a point can be at the end of a sentence. Nevertheless, the experience in the work with documents confirms the efficiency of such testing. Only after analysis of points, which are not sentence markers, we can divide the paragraph into sentences and the further analysis is performed within one sentence.

Description of the algorithm of a morphological analyzer

The work of a morphological analyzer is as follows. At its input, we have an array of words, punctuation marks and numbers marked from the input text at the stage of lexical analysis, with lexical characteristics [11]. For every word, the analyzer performs the search for words in the dictionary loaded into memory. All the stems the word being analyzed can begin with are looked for. If a stem in turn satisfies this condition, a line containing all possible affixes for this stem is extracted from the dictionary of affixes. Each affix from this line is added by turns to the stem and the result is compared with the word being analyzed. In case of their exact coincidence, a new record is introduced into the list of the search results: according to the ordinal number of the affix in the line of affixes, variable morphological parameters of the word are determined (for example, for a noun – the number and case), and by the lexical information of this stem its constant parameters (noun, verb, adj…) are determined. If as a result of such search not a single successful variant if

found, the user is requested to enter a new stem into the dictionary. In case he refuses to do it, performance of the morphological analysis is stopped. If the new word is introduced into the dictionary, the procedure of searching is repeated. At the output of the morphological analyzer, we have a list of lemmas (a normal form of the word) + affix + morphological characteristics of the word (part of speech, case, number).

Thus, the results of the morphological analysis in total are presented in the form of a dynamic array. The number of its elements is equal to the number of its lexemes in the sentence. The elements of the array are other arrays each of which retains all possible interpretations of its lexemes homonyms. A dictionary of word stems, a dictionary of geographic names, a dictionary of names, a dictionary of affix conjunctions are used as initial lexical materials.

Algorithm is automated filling of the ontology of factographic retrieval system

The section describes the application of morphological parser of the Kazakh language to automated filling of the ontology of factographic retrieval system. The method of automated filling of the ontology is proposed in [12]. This algorithm allows to extract key words/ phrases from the text corpus of homogeneous subjects. The extracted key words are used further as possible meaning of attributes of matters described in the created ontology of the subject field designed for organization of factographic retrieval. The text preliminarily marked up with the help of a specially developed morphological analyzer is used as input data. To extract semantically related key words/ phrases, the method of random walks is used in the algorithm. A trained neural network with a hidden layer is used for the set of these phrases with the aim to assign a concrete phrase to the definite attribute of the matter described in the text. Owing to the neural network operation, ontology for a concrete document on the semantically related word pairs set is constructed and then, on the basis of the obtained ontology, factographic retrieval is organized.

Experiments

The biography of Kazakh poetess Fariza Ongarsynova is taken as an example of a complete cycle of the algorithm work. At the first stage, the text of biography is marked up by parts of speech. For example: *Фариза Оңғарсынқызы Оңғарсынова - қазақ ақыны, халық жазушысы, журналист. 1939 жылы 5 желтоқсанда Гурьев (қазіргі Атырау) облысы, Новобогат ауданына қарасты Манаш ауылында туған.* In this form, the document is introduced in the database MongoDB which is oriented to store collections of JSON documents.

Then, using the random walk method, key words and phrases are extracted, part of them are presented below: *Фариза Оңғарсынқызы Оңғарсынова, ақын, халық жазушы, 1939 жылы 5 желтоқсанда, Гурьев облысы, Атырау, Новобогат ауданы, Манаш ауылы, etc.*

And in the last stage, the neural network places the data on the descriptors.

"name": "*фариза оңғарсынқызы оңғарсынова*"

"position": "*ақын*", "*жазушы*"

"date_of_birth": "*1939 жыл 5 желтоқсан*"

"date_of_death": "2014 жылдың 23 қаңтар"

Conclusion

The work proposes the technology of automated filling of the ontology of factographic retrieval system. The proposed technology allows to markup parts of speech in the text. The authors cover the progress of the work on supplying the corpus of the Kazakh language texts with scientific framework. While making morphological marking they considered the features of the Kazakh word form change, and created the models of variable nominal and verbal word forms and the lists of their word changing affixes. In the future we plan to continue research to develop modules of syntactic and semantic analysers that expand opportunities for the filling of the ontology of factographic retrieval system.

References

1. Makhmudov M.: Systems of automatic recycling of Turkic text on lexical and morphological level, Elm, 114 p. Baku (1991) (In Russian)
2. Migalkin V.V.: Modeling of the Yakut language spelling and development a set of programs to check the spelling of Yakut texts in Windows environment, Author. diss.... Ph.D., Yakutsk (2005) (In Russian)
3. Sadyqov T.: Problems of modeling of Turkic morphology: an aspect of causing Kyrgyz nominal inflectional forms, 119 p. Publishing House of the "Ilim" (1987) (In Russian)
4. Sirazitdinov Z.A.: Modeling grammar of Bashkir language. Inflectional system. 160 p. Ufa (2006) (In Russian)
5. Sirazitdinov Z.A.: On the modeling of inflectional system agglutinative language pair combinations (for example, the Bashkir language) / Actual problems of modern Mongolian and Altaic. Proceedings of the International Scientific Conference. Elista, 2014. pp 139-143. (In Russian)
6. Altenbek G., Wang Xiao-long: Kazakh Segmentation System of Inflectional Affixes. In: Joint Conference on Chinese Language Processing, pp.183-190 (2010)
7. Zafer H.R., Tilki B., Kurt A., Kara M.: Two-level description of Kazakh morphology. In: Proceedings of the first International Conference on Foreign Language teaching and Applied Linguistics, FLTAL 2011, Sarajevo (May 2011).
8. Sharipbaev A.A.: Intelligent morphological analyzer, based on semantic networks: Conference proceedings Open Semantic Technologies for Intelligent Systems (2012)
9. Bekmanova G.T.: Some approaches to the problems of automatic word changes and morphological analysis in the Kazakh language. In: Bulletin of the East Kazakhstan State Technical University Named by D. Serikbayev, №1, pp. 192-197, Ust-Kamenogorsk (2009) (In Russian)
10. Zhubanov A.H.: Basic principles of formalization of the Kazakh text content, 250 p. Almaty (2002) (In Russian)

УДК 81.374:811.512.161:811.512.154

PARSING AND ANNOTATION OF TURKISH-KYRGYZ DICTIONARY

Kadyr Momunaliev, *Kyrgyz-Turkish Manas University kadyr.momunaliev@gmai.com*

The case study described in this article is the first milestone on the way toward a full featured Text Encoding Initiative P5 annotation standard. The paper outlines parsing and annotating workflow to obtain initial XML-based structure of Turkish-Kyrgyz dictionary. Typography-based parsing techniques are implemented in procedural programming language environment; corresponding workflow charts are presented in form of pseudo code and block schemas. Resulted XML dictionary bases are verified and applied in desktop and web e-dictionary implementations. It is proposed that such kind of explicitly structured data representation is easier to manipulate and use as a basement for further deeper lexicographic annotations.

Keywords: dictionary data, structured data, unstructured data, XML, human-computer, typography, semantics, syntax, parsing

ПАРСИРОВАНИЕ И АННОТИРОВАНИЕ ТУРЕЦКО-КЫРГЫЗСКОГО СЛОВАРЯ

Момуналиев Кадыр Замирович, *Кыргызско-Турецкий Университет «Манас» kadyr.momunaliev@gmai.com*

Данное тематическое исследование представляет собой один из пройденных этапов на пути к достижению полноценного стандарта аннотирования Text Encoding Initiative P5.