Article

JRIE

Morphological Analyzer in Corpus Linguistics as a Mathematical Model

Journal of Research in
International Education
2023, Vol. 22(2) 146–149
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14752409276723553
https://www.jrie.org
SAGE

E F Sharipova
Student, Gulistan State Pedagogical Institute, Republic of Uzbekistan

Abstract

This article examines the issues of corpus linguistics from the preface to a summary of the situation with his wife. That is, the investment of any socio-humanitarian disciplines in mathematical models, in short, molding or digitization, forms an innovative environment in the world of science. In particular, the fundamental basis of science has been formed in Uzbek linguistics, which further increases the potential of computational linguistics, which is widely covered on the Internet today, that is, it is recognized as a novelty.

Keywords

morphological analyzer, automatic analysis, grammatical analysis, morphological analysis, syntactic analysis, semantic analysis, lemma, stem, tag, prosodic tag, anaphoric tag, semantic tag, morphological tag, syntactic tag.

Introduction

Automatic text analysis consists of several complex operations: the computer analyzes the text in natural language according to a given algorithm, during the automatic analysis, the computer forms a lexical-morphological, syntactic, semantic representation of the text in its language. The process of automatic text analysis and synthesis is considered one of the main tasks of computational linguistics [Öztürel, et al., 2019; Sadykov, 2018; Tantug, et al., 2006]. This task stems from the creation of linguistic models of artificial intelligence and the practical needs of mankind (for example, automatic machine translation). [Sharipov, 2022]

Automatic text analysis

Automatic text analysis consists of several stages:
1) graphematic analysis: determining the boundaries of a word, sentence, paragraph and other text element (for example, a fragment of newspaper text;
2) morphological analysis: determining the shape of the head of the word used in the text, as well as the morphological features of this word;
3) syntactic analysis: determining the grammatical structure of a sentence in the text;
4) semantic analysis: differentiation of the meanings of phrases [Bolshakova, 2011].

One of the above types of analysis is graphematic tokenization (eng. the token. - a word, phrase, or any significant element in the text). While the formal signal space (separation) between words as a determinant of the boundary of text elements, the limit of a sentence and its parts are distinguished by capital letters and punctuation marks, the substantial parts of the text are paragraphs.

Corresponding author:
E’zoza F. Sharipova, Student, Gulistan State Pedagogical Institute, 120100, 4th-microregion, Bo‘ston mahalla, campus of GulSPI
Email: fazliddindpr@mail.ru

But the formal method is not always useful when determining the boundary of a word. For example, there is no formal word boundary in Chinese. Even the fact that in most European languages, several words separated by a space (division) are equivalent to a single lexeme shows that graphemic analysis using this method does not yield results.

Looking at the second analysis, i.e. morphological analysis of the text, the head form of the word form used in it, the Lemma, is revealed, and the grammatical meanings of the word in this context are highlighted. For example, a category for a noun, person, number, conjugation, person, number, tense, etc. for a verb... Each word, word form, or word used in the text is called a use case. In inflectional languages, the core (base) in morphological analysis may not coincide with the base in traditional grammar. The determination of the base in morphological analysis is called stemming. Stem is a part of a word form without syntactic formers, with the grammatical basis of some Turkic kernels that have undergone the phenomenon of inflection, in the process of morphoanalysis – there are cases of STEM incompatibility. For example, the base of the verb “sana” (count), formed from the base of a “san”, is a number, but this phrase is the core of the form. Here, the word-former is extracted from the core and remains san stem. But in agglutinative languages, including Uzbek, the Lemma is usually equivalent to STEM, and the state of flexion has little effect on the process of morphoanalysis. [Khamroyeva, 2020]

In morphoanalysis, only the lemmatization process itself is missing, and stemming also plays a role. In particular, stem will be a key factor in improving the quality of internet search, which means that search will be based on STEM rather than lemma. [Toldova & Bonch-Osmolovskaya, 2011] Therefore, morphological analysis does not end with computer understanding of the word form, that is, computer differentiation of the sequence of characters will not be enough. The analysis does not end with the process of finding the difference between the word form and all its members in the paradigm. The main task is to find both categorical specifics (tagging) and morphological signs of word formation, and suffixes expressing grammatical meaning (computer models), it is advisable to form both formers of a certain lexeme and its paradigm member. To determine the state of lexico-morphological homonymy in words, the word relies on both the syntactic and semantic nature of the sentence so that it can correctly attach morphological features to the form. Thus, it is assumed that this will be done automatically and with a human factor. When such processes are automatically tagged, rule modules are launched [Öztürel, et al., 2019; Sadykov, 2018; Tantug, et al., 2006].

As you can see, thanks to this set of rules, that is, determining the order of groups of words and parts of a sentence in a sentence, it becomes possible to distinguish homonymy.

Khamroyeva's monograph lists three ways to build a morphological analyzer: [Khamroyeva, 60]
1) Dictionary-based analyzer;
2) an analyzer based on a set of grammatical patterns without a dictionary;
3) an analyzer based on a dictionary and grammatical rules.

At the same time, we see that the phenomenon of synharmonism among the languages belonging to the Altai family is based on the rules and evidence of the relatively widespread Kyrgyz language and is implemented in four stages:
1) dividing the words of the introductory text into grammatical forms;
2) lemmatization of the word, i.e. from the dictionary of fundamentals, finding the lexeme form of the word, definition;
3) clear separation of the chain of adverbs forming the syntactic form;
4) determination of the morphological feature of each affix.

We see how the researcher compares the facts of the Uzbek and Kyrgyz languages to explain the essence of the morphological analyzer. For example, in the word-formation analysis of baldar (children), the morphological analyzer must determine whether these forms are formed from a child base with the addition of a plural suffix.

The phenomenon of synharmonism in the Kyrgyz language as a result of the addition of the suffix -f to the base led to the loss of the last letter in the base of the ball. In this case, the analyzer should lemm the word, automatically determining the basis before the suffix. This will launch a basic dictionary search. When parsing a book word form, the morphological analyzer must determine whether the possessive form is attached to the book word-im: the book is defined as the basis. In the Kyrgyz language, the word “kitebim” of the possessive form, when the possessive suffix – “im” is added to the basis of “kitep”, under the influence of the suffix-unstressed “P” turns into a sonorous sound “B”. In this case, the dictionary of the basics is equipped with additional rules. The absence of a large number of phonetic changes in the Uzbek language can shorten the stage of algorithms in the process of morphological analysis. But in the case when a phonetic change occurs, the process of determining the core of a word that is undergoing a change, as in the Kyrgyz language, is performed according to an algorithm of 2-3 steps. [Khamroyeva, 60-61]

It follows from this that among the principles of developing linguistic support for the construction of a morphological analyzer of the Uzbek language, the main task is to form this block of rules – a linguistic module. It is also important to create a morphological database in a morphological analyzer. This implies a wide range of data that meets the needs of the user. This data is sorted and stored as a table managed by a database management system. Currently, there are many systems designed to work with databases: SQL, MySQL, Oracle, Accss. It is always difficult to work with large amounts of data, although each system has its own advantage.

To implement morphological analysis, in order to test the natural language and verify the proposed algorithm, the Embarcadero RAD Studio test program was created, this analyzer used the Access database system. This analyzer consists of three database tables (base, complement, and word category), as well as the relationships between these tables.

Researcher Sh.Khamroeva notes that it is difficult to create an ideal morphological analyzer. Nevertheless, the implementation of the following algorithm for constructing a morphological analyzer can become the basis for creating an analyzer that performs a relatively accurate analysis:
1) building a frame model of the grammatical form of a word. This includes creating a table of suffixes related to the category to correctly identify the category of the auxiliary morpheme.;
2) development of algorithms designed to reduce the number of accesses to computer memory when compiling the analyzer;
3) Developing a testing program, experimenting with the results in it [Khamroyeva, 63].

Our other researcher, K.Shadmanova, argues that one of the theoretical foundations of system programming is the theory of formal language. This theory with its mathematical concepts will seem complicated to an ordinary programmer, but a system application is created on the basis of such a theory. Any programming interface is based on formalism, and formalism is based on the theory of formal language. In general, for a sentence constructed in an arbitrary language (form or natural) to make sense, it must satisfy the following conditions:
1) words must be written in accordance with the requirements of the alphabet (morphology, vocabulary);
2) avoid grammatical errors in the construction of a sentence (grammatical, syntactic);
3) the correctness of the meaningful construction of the sentence (semantics).

The development of morphological analyzers of Turkic languages began in the 60s of the last century, [Suleymanov, et al., 2003, 220] the main feature of the early analyzers was that they were not designed specifically for one language: one morphoanalyzer could be adapted to another language. Since morphotactic rules are attached to the program, the system of rules is supplemented by the vocabulary of the language, as well as affixal morphemes, it becomes possible to use it as an analyzer of another language; there is no need to write separate codes to create an analyzer of another language. Since then and up to the present time, technologies have changed, universal morphological analyzers of Turkic languages have been created, the volume of dictionaries and the speed of information processing have increased. Despite the fact that more than 50 years have passed since the movement to develop a morphoanalyzer of Turkic languages began, the growth of this field is still at different levels in all Turkic languages. The works on Tatar, [Suleymanov, et al., 2003, 220] Bashkir, Kazakh, Chuvash, Turkish, Khakass languages, as well as on the universal morphological analyzer confirm our opinion. [Suleymanov, et al., 2019, 39-40; Marchuk, 2007, 65; Suleymanov, et al, 2003, 220]

The literature indicates such types of automatic analysis as stemming, dictionary analysis of word formation, analysis based on a logical approach, tabular analysis, analysis without a dictionary. Experts note that automatic morphological analysis consists of such basic blocks as stemmatization, lemmatization, and grammatization. [Suleymanov, et al., 2019, 39-40; Marchuk, 2007, 65]

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Большакова Е. И. И др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебная пособие. – Москва: МИЭМ, 2011. 272 с. (in Russian)

Сулейманов Д. Ш., Хусаинов А. Ф. (Ред.) Формальные модели и программные инструменты компьютерной обработки татарского языка: монография. Казань: Институт прикладной семиотики Академии наук Республики Татарстан, 2019. 260 с. (in Russian)

Марчук Ю. Н. Компьютерная лингвистика: учебное пособие. – Москва: АСТ Восток-Запад, 2007. 317 с. (in Russian)

Сулейманов Д. Ш., Гатиатуллин А. Р. Структурно-функциональная компьютерная модель татарских морфем. – Казань: Издательство “Фэн” Академии наук Республики Татарстан, 2003. 220 с. (in Russian)

Толдова С. Ю., Бонч-Осмоловская А. А. Автоматический морфологический. Анализ. – Москва: Фонд знаний «Ломоносов», 2011. URL: https://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127430

Khamroyeva Sh. O’zbek tili morfologik analizatorining lingivistik ta’minoti. Toshkent: ClobeEdit, 2020. – B.50.

Sharipov F. O’zbek zamonaviy morfologik nazariyasining shakllanishi, taraqqiyot tendentsiyalari va istiqbollari. Fil.fan.dok.diss. –Guliston, 2022. 255 b.

Öztürel A., Kayadelen T., Demirşahin I. A Syntactically Expressive Morphological Analyzer for Turkish // FSMNLP 2019 – 14th International Conference on Finite-State Methods and Natural Language Processing, 2019. URL: https://doi.org/10.18653/v1/w19-3110

Садыков Т., Кочконбаева Б. Об оптимизации алгоритма морфологического анализа // Шестая Международная конференция по компьютерной обработке тюркских языков «Turklang-2018». – Ташкент, 2018. 320 с. (in Russian)

Tantug C., Adali E., Oflazer K. Computer Analysis of the Turkmen Language Morphology // 5th International Conference on NLP “FinTAL-2006”. – Turku, Finland, 2006. – pp. 186-193.

Auhtor biography

E’zoza F Sharipova, Student, Gulistan State Pedagogical Institute, Republic of Uzbekistan.