Implementation of the Removing Homonymy by Collocation System

Anastasiia Khluieva, Zoia Kochuieva, Natalia Borysova

National Technical University "Kharkiv Polytechnic Institute"2, Kyrpychova str., 61002, Kharkiv, Ukraine

Abstract

This article describes implementation of the removing homonymy by collocation system method in Ukrainian language, which can be implemented on Python programming language. It also includes the relevance of removing homonymy phenomenon and difficulties associated with it. A study of various methods of removing homonymy was carried out and conclusions are mentioned in this work. In article there is an algorithm of the removing of homonymy, particularly homoforms, by collocation system method.

Keywords 1

Homonymy, methods of removing homonymy, collocations, corpora

1. Introduction

Modern intellectual systems do not process texts in natural languages with sufficient quality due to the presence of such a linguistic phenomenon as homonymy. This concept is marked by ambiguity of approaches to its study and, in fact, interpretations. This has caused a relentless interest on the part of linguists, because the phenomenon still remains relevant to his research. Homonymy is a difficult phenomenon, it has many aspects that require comprehensive analysis.

The main purpose of its research is usually classification; sources of origin of homonyms and their distinction from polysemous words; issues of interlingual homonymy, as well as the removal of homonymy in the automation of translations, etc.

Thus, the purpose of this work is to develop a system for removing homonymy using the method of collocations.

To achieve this goal, we have the concept of homonymy as a linguistic phenomenon, its types and problems of origin, as well as possible methods of removing homonymy. The subject of the study is the removal of homonymy by using the method of collocations. To solve this problem, the following tasks were formulated:

1. Performing an analytical review of the literature on the topic of homonymy, its types, problems that arise for its removal, and methods for eliminating this problem.
2. Development of an algorithm for removing homonymy by collocation.

2. Homonymy Phenomenon

Homonymy is a synchronous phenomenon in terminology, which is based on the absence of common sem in the meanings of the same terms of expression of terms and commonly used words, terms of one or more related or unrelated areas of knowledge and human activity. Depending on the formal, more precisely, on the grammatical (morphological) and semantic, relations between homonymous words, there are several types of homonymy: lexical (sound coincidence of different in meaning linguistic units belonging to the same part of speech), grammatical (sound coincidence in

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine

EMAIL: nastiahlueva@gmail.com (A. Khluieva) ; kochueva@kochuev.com (Z. Kochuieva) ; borysova.n.v@gmail.com (N. Borysova)
ORCID: 0000-0002-4300-3370 (Z. Kochuieva); 0000-0002-8834-2536 (N. Borysova)

separate grammatical forms of language units of different meanings), word-forming (sound coincidence of morphemes of different word-formative meanings), syntactic (sound coincidence of different syntactic constructions), phonetic (a sound coincidence of language units of different meanings with different spellings), graphic (a graphic coincidence of language units with different pronunciations). Among the grammatical homonymy there are homoforms, homographs, homophones. Homoforms are morphological homonyms that are distinguished on the basis of sound coincidence and the same spelling of word forms belonging to different lexical and grammatical classes or different forms of the same word. Precisely this type is an object of our study.

3. Removing Homonymy Methods

Historically, almost all methods of removing homonymy are divided into two groups:

1. Methods based on rules. In turn, are divided into:

- Methods with manual entry of rules.

An illustrative example of a method with automatic rule generation is the method of the American linguist Eric Brill. Transformation rules are a set of "old tag, new tag, condition", and the application of the rule is to replace the old tag with a new one when the specified condition. The disadvantage of this method is the decrease in the increase in accuracy with increasing number of rules, which, however, is fully consistent with the Pareto principle: "80% of the effort provides 20% of the result." At the same time, the principle works in the opposite direction: performing only one initialization step is enough to achieve high accuracy of homonymy removal.

2. Methods with automatic rule generation.

- Methods based on statistics.

Statistical methods of removing homonymy allow us to calculate the probability of each possible variant that occurs on the basis of statistics: if in any context the noun occurs more often than the union, the homonym found in the same context will be more likely to be a noun than a union (if these options allowed by the dictionary). The disadvantages of probabilistic methods are the duration of formation and marking of the body of texts and low accuracy of analysis, caused by the free order of words in inflected languages. Thus, there is a need to create a method of removing ambiguity for the homoform of different parts of speech, which does not require a large number of rules or a body of manually marked texts.

Therefore, there is a need to develop a hybrid method that uses both rules and information from texts published on the Internet, which does not require repetition of the parsing procedure in the case of homonymy.

4. Collocation System Method

To get started, we need to address the issue of collocations. As it’s known, collocations are compounds of words, the probability of using which together is greater than the probability of using these words separately from each other. This issue of collocation research is one of the leading ones in applied linguistics. Precisely because these compounds are usually stable, they need a certain form of words to match each other, which can help with the removal of homonymy. Based on the results of the analysis of sources on the problem of homonymy, namely the emergence of incomplete grammatical homonyms, among homonymous pairs of words homonymy with the noun most often occurs, so we take collocations with nouns to remove homonymy. From the works of Ukrainian scientist Bobkova T.V. [1], we found that collocations with nouns in the Ukrainian language are characterized by the following part-of-speech combinations:

1. adjective + noun,
2. preposition + noun,
3. verb + noun,
4. noun + noun.

Among the above types of collocations in the language, collocations such as “adjective + noun” (1) (ukr. “широке поле”) and “preposition + noun” (ukr. “після дати”) (2) are most often used, and we will take such collocations for our research.

To use type 1 collocations, it is necessary to find homonymous pairs, to investigate the context of these words. To get the most commonly used context with these words, we will use the Sketch Engine [2] platform, where you can find the context for a word from a text in the Ukrainian language with a volume of more than 2 billion words.

To use type 2 collocations, we will refer to the frequency dictionary of collocations and delete collocations with a frequency of more than 30 cases per 1 million words.

The base of collocation is 821 collocations (615 collocations of preposition and noun and 206 collocations of adjective and noun). In our study, we will not reduce the word to the lemma, because the study of homoforms requires exactly the form in which the problem of homonymy arises. Therefore, the collocation database will use the form of the word in which difficulties arise, with the appropriate context for comparison directly with the form of the word used in the sentence.

Hence, on the basis of the revealed sequences we created particular algorithm of work:

1. The user enters text in the interface window, then text goes through the tagging process.
2. The program checks whether the preposition with context is contained in our database. Its tag changes to the one specified in the database as long as the match is found.
3. The program checks whether the noun is contained in the context of the adjective from the database. If the match is found, its tag changes to the one specified in the database.
4. As a result, the program produces text with assigned parts of speech to each word.
5. The user receives standard tagged text provided there is no matches are found in the databases.

Given method and algorithm can be easily extended for usage of noun and not only, and continue to increase the base of collocations for Ukrainian language.

5. References

[1] Sketch Engine concordance for Ukrainian language based on Ukrainian Web 2014 corpus. URL: https://auth.sketchengine.eu/#login?next=https%3A2%2Fapp.sketchengine.eu%2F%23dashboard%3Fcorpname%3Dpreloaded 252Fuatenten14

[2] Linguistic portal Mova.info Dictionary Of Ukrainian Prepositional Collocations. URL: http://www.mova.info/Page.aspx?l1=6