IMPORTANCE OF STATISTICAL METHODS IN DETERMINE THE HOMONYMY

Axmedova Xolisxon Ilxomovna
PhD student
Tashkent state university Uzbek language
and literature named after Alisher Navai

Abstract: The problem of automatic processing of natural language remains relevant for more than half a century. One of the important problems in the field of NLP is the creation of a semantic analyzer, which in turn goes through a series of steps. Determining homonymy is important in the semantic analysis of sentences. Statistical methods are also used to determine homonymy. The frequency method is used to determine homonymy between grammatically similar word groups. This method involves extracting homonym classification parameters.

Keywords: NLP, semantic analyzer, hoonomy , method based on statistical data, classification parameters, Hidden Markov model.

Аннотация: Проблема автоматической обработки естественного языка остается актуальной уже более полувека. Одной из важных задач в области НЛП является создание семантического анализатора, который в свою очередь проходит ряд этапов. Определение омонимии важно при смысловом анализе предложений. Статистические методы также используются для определения омонимии. Частотный метод используется для определения омонимии между грамматически сходными группами слов. Этот метод включает в себя извлечение параметров классификации омонимов.

Ключевые слова: NLP, семантический анализатор, хуномия, метод на основе статистических данных, параметры классификации, Скрытая марковская модель.

Natural Language Processing (NLP) – a general area of artificial intelligence and mathematicallinguistics, which studies the problems of computer analysis and synthesis of sentences in natural languages. Solving this problem means creating a more convenient form of interaction between a person and a computer. The problem of automatic processing of natural language remains relevant for more than half a century. The complexity of the problem and the lack of a clear idea indicate the difficulty of ways to solve it. Linguistic analyzers are particularly important as tools for automatic processing of sentences. Linguistic analyzers are divided into morphological, syntactic and semantic analyzers. They, are divided into several groups.

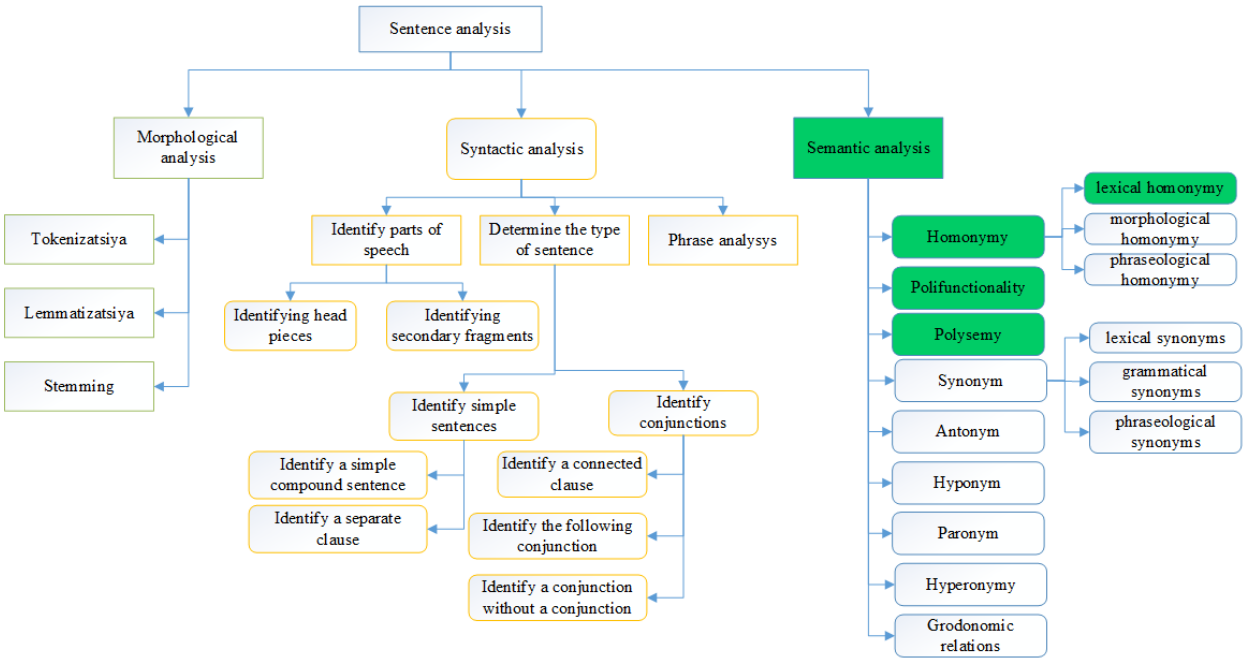

Figure 1. Hierarchy of linguistic analyzers and their elements

Each linguistic analyzer described in the hierarchy in Figure 1 has its own task. A morphological analyzer is a set of algorithms that compare individual words/word forms in the dictionary and determine the grammatical properties of words.A syntactic analyzer is an information system that syntactically analyzes an input sentence based on the characteristics of the input language.

A semantic analyzer is an information system that determines the content of the sentence through the meanings of the words in the sentence. Semantic search is performed through semantic analysis. The better the semantic analysis is developed, the more effective the search results will be. Implementation of semantic analysis directly depends on linguistic resources. Lexical resources include dictionaries, thesauruses and ontologies. Semantic analysis also has its own elements, which require a separate study. One of the important elements of semantic analysis is homonymy. Homonymy detection is interpreted differently in different natural languages. In world computer linguistics, 3 methods are mainly used in the semantic analysis of sentences:

- Rule-based methods;
- Method based on statistical data;
- A method based on machine learning.

These methods are used differently in different languages. Homonymy Baltabayeva J.K. and Sulaymanova J.N (Kazakh language) [2019], Ch.A. Davlyatova (Tajik language) [2017], V.V. Kukanova (Bashkir language) [2014], H. Heydarova (Azerbaijani language) [2017] V.V. Kukanova (Kalmyk language) [2011] has been the subject of research by Turkologist scholars.These Turkic scientists rely on the theory of homonymy developed by A.I.Smirnisky, V.V.Vinogradov, O.S.Akhmanova and other linguists. The results of the research of Turkic scientists show that the methods mentioned above are important for determining homonymy.In this article, we will try to highlight the use of statistical methods and its importance in distinguishing homonyms in the Uzbek language.

The method based on statistical data is used to solve the problem by classifying the grammatical parameters of words. These parameters are chosen differently in different natural languages. For example, when determining morphological homonymy in Russian, parameters such as word group, gender (genus) of the word, singular or plural form, lemma, lemma and word group, only lemma, and homonymy are distinguished [1]. The issue of removing morphological homonymy (snyatie homonymy) in the SemSin system of the Russian language is solved. In Russian, this term is called "snyatie homonym", because if the identified elements of the semantic analyzer are removed from the text, the analysis of the remaining elements increases the speed of the system, ensures the accurate and correct operation of the algorithms. Classification parameters for the Uzbek language can be divided as follows

Having studied foreign experiences in depth, we use rule-based, statistical data-based and machine learning-based methods to distinguish homonyms in the Uzbek language. When distinguishing homonyms in the Uzbek language, we divided them into groups such as homonyms within one word group, two word groups, three word groups, and four word groups according to their occurrence within word groups. We used the rule-based method to determine homonym within grammatically dissimilar word groups. This is covered in detail in articles [4,5,6].

The method based on statistical data} is used to solve the problem by classifying the grammatical parameters of words. These parameters are chosen differently in different natural languages. For example, when determining morphological homonym in Russian, parameters such as word group, gender (genus) of the word, singular or plural form, lemma, lemma and word group, only lemma, and homonym are distinguished [1]. The issue of removing morphological homonymy (snyatie omonimii) in the SemSin system of the Russian language is solved. In Russian, this term is called "snyatie omonimii", because if the identified elements of the semantic analyzer are removed from the text, the analysis of the remaining elements increases the speed of the system, ensures the accurate and correct operation of the algorithms. Classification parameters for the Uzbek language can be divided as follows in figure 2.

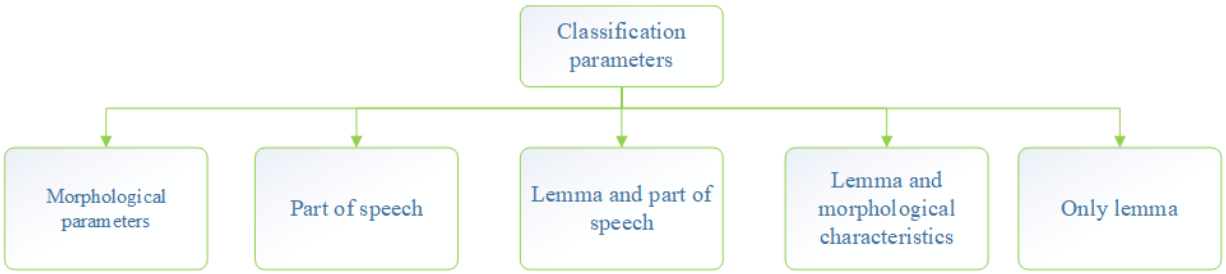

Figure 2.Classification parameters

If the level of accuracy of the semantic analysis of the word is not high, the classification parameters can be expanded. Statistical methods can be used using extracted parameters. The main task of this method is to divide the content of the context into n-grams, that is, the combinations of the input word in the context are determined and evaluated using evaluation methods [1]. In some statistical methods, a decision is made using the word and its suffixes, while in some, it is concluded using the semantic valence of the word in the context. It follows that methods based on statistical data are divided into several groups according to decision-making parameters(fig. 3).

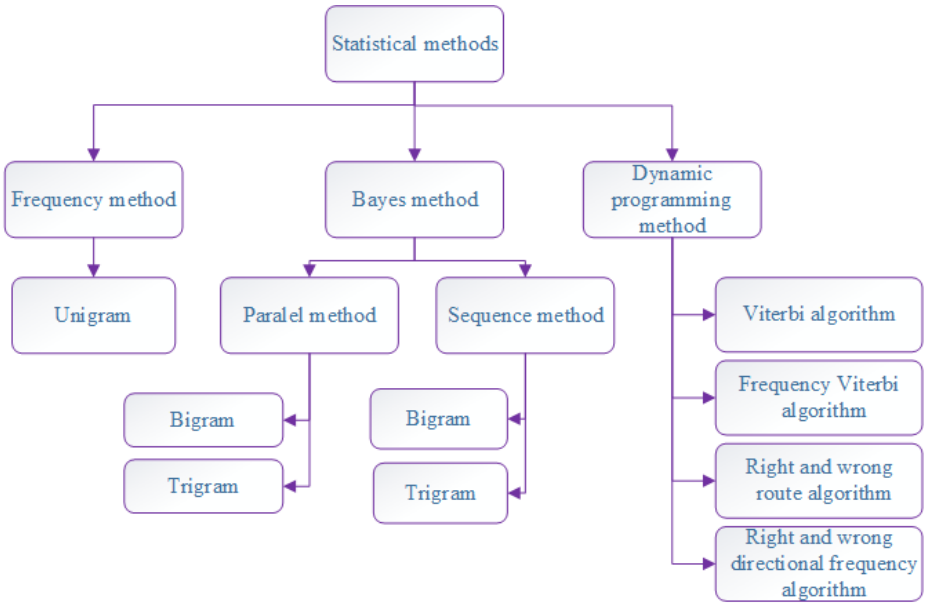

Figure 3. Hierarchy of statistical methods

As shown in Figure 1, the Frequentest method is semantically analyzed through the unigram, that is, the word itself, while the Bayesian method uses bigrams and trigrams of the word. Dynamic programming method uses n-grams. Context tokenization and tagging processes are carried out in the evaluation by these methods. Using statistical methods, we will consider the process of differentiating the meanings of homonyms within different word groups. For example, let's take a sequence of identifying homonym between noun or adjective word groups using the frequency method.

Homonyms within the *noun ∨ adjective* group are classified according to the following parameters

- Part of speech;
- Stem and lemma;
- Only lemma
- Only stem

Given a sentence with a homonym from the noun ∨ adjective group.

Isitma, yo'tal, burun bitishi va bosh og'rig'i, bularning hammasi gripp **alomatlari** *hisoblanadi.*

In this sentence, the word "alomat" is a homonym and has the following meanings.

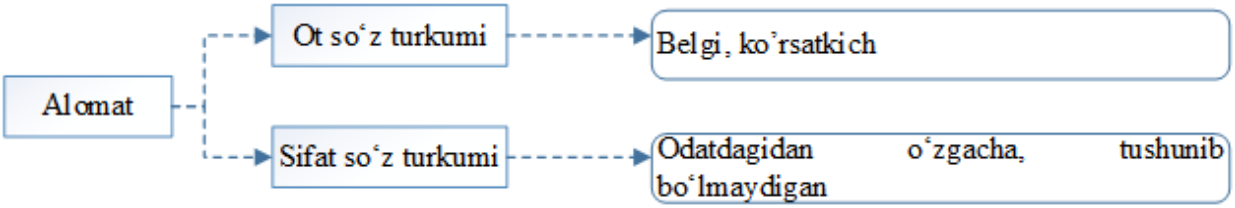

Figure-4. Definitions of alomat word

When classifying the word "alomat" according to the above parameters, we use the data of the national corpus of the Uzbek language. A total of 6,823 pieces of information were found

when the word “alomat” was searched through the Uzschoolcorpora.uz site, 100 of which were analyzed

79 of them are nouns

21 are adjectives

The analyzed 100 pieces of information were divided into core and appendices and the following results were determined.(fig.5)

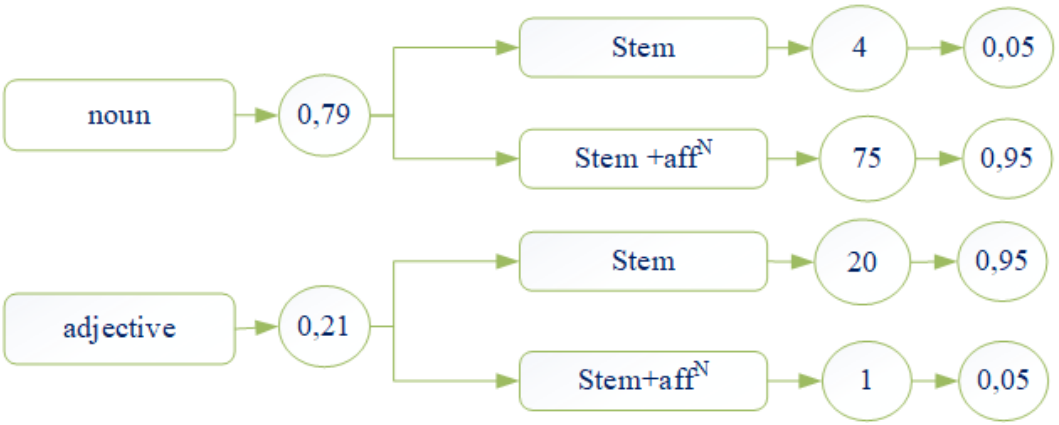

Figure 5.Statistical information

Based on the obtained statistical data, we also divide the homonym in the given sentence into stems and suffixes (fig.6)

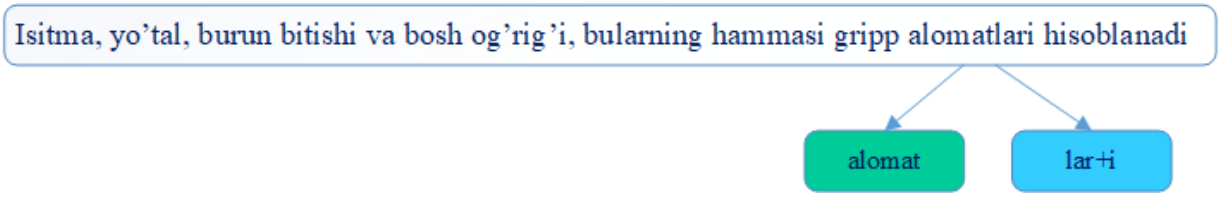

Figure 6. A sentence with the word “alomat”

So, based on the above statistics, the following decision is made for the word “alomat” in this sentence.(fig.7)

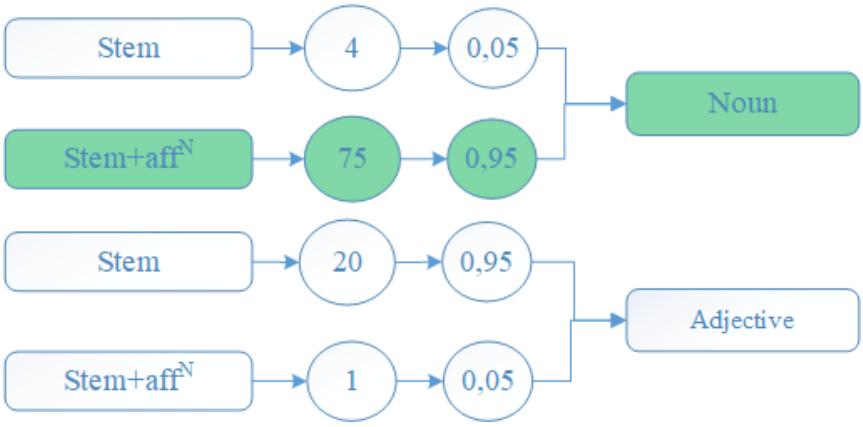

Figure 7. Homonym identification using classification parameters

The information in this chart shows that the word "alomat" in the given sentence is a homonym of the noun group with a 95% probability, and it means a sign, an indicator. In this way, homonym can be determined by frequency method. It can be seen that in order to use the frequency method, it is required to determine
the classification parameters for each of the homonyms within different word groups and perform statistical calculations based on them. Another common probabilistic approach is an algorithm based on the use of a Hidden Markov Model (HMM). The main idea of the algorithm is to choose a Grammar tag that maximizes the value of the following function for each word in the sentence:

$$P(word\ |tag)*P(tag\ |previous\ n\ tag\ )$$

Here, P(tag |previous n tag)- conditional probability (estimated in the corpus), the probability of occurrence of the current tag with n predefined tags, P(word |tag)- conditional probability (calculated using corpus data) is a tag determined based on Grammatical properties of the word. Although HMM has complex computations, it has various simplifications in practice. Distinguishes word meanings with 96 % accuracy for English grammar. Applying this model to Russian may be difficult compared to English, requiring very large corpora given the richness of word formation and word variation in Russian.

In Tatar, as in other groups of agglutinative Turkic languages, morpheme is the most important meaningful unit of linguistics, carrying both semantic and syntactic information. TuganTel - an algorithm for determining morphological polysemy in the corpus of the Tatar language, a database has been developed.

Gataullin Ramil Raisovich discussed models and algorithms for removing homonym and polysemy in the Tatar language in his candidate thesis. The use of Markov models in distinguishing homonym and polysemy was also recommended in the Tatar language [2]. Having studied foreign experiences, it can be concluded that it is appropriate to use statistical methods to identify homonyms of different word groups in the Uzbek language.

References:

1. Rysakov S.V. Klyshinsky E.S.: Statisticheskie metody snyatiya homonym// Novye informatsionnye tehnologii v avtomatizirovannyx sistemax. – 2015 Springer, Heidelberg (2016). doi{10.10007/1234567890
2. Gataullin Ramil Raisovich.: Method, model and programmatic instrumental resolution of multivariateness in text// diss-VAK RF-05.13.11 kand.nauk.-(2019)
3. Chirag Goyal-June 23, 2021: https://www.analyticsvidhya.com/blog/2021/06/part-9-step-by-step-guide-to-master-nlp-semantic-analysis/
4. Elov B.B, Axmedova X.I.: Uchta so‘z turkumi doirasidagi omonimiyani farqlovchi biznes jarayonni modellashtirish//O’zbekiston respublikasi innovatsion rivojlanish vazirligining, ilm-fan va innovasion rivojlanish ilmiy jurnal 2022 / 1
5. Sh.K. Gulyamova, X.I. Akhmedova. Linguistic basis of homonyms, mathematical model and algorithms (in the framework of nouns and verbs, adjectives and verbs) // KoqonDPI. Scientific reports, 20211`1
6. Axmedova X.I. O‘zbek tilidagi shakldosh so‘zlarni semantik tahlil qilish // ‘‘O‘zbekistonning umidli yoshlari’’ Mavzusidagi 7-son Republika ilmiy talabalar, magister yosh tadqiqotchilar va mustakil izlanuvchilar uchun onlayn anjuman sining materiallari heat. - Toshkent: tadqiqot, 2021.