Methods of eliminating homonymy within different, grammatically similar word groups

Sh. S. Sirojiddinov[1*], *B. B.* Elov[1], and *X. I.* Axmedova[1]

[1]Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Tashkent, Uzbekistan

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 number of steps. Determining homonymy is important in the semantic analysis of sentences. A method based on rules, a method based on statistical data, and methods based on machine learning are also used to determine homonymy. Statistical methods are mainly used to determine homonymy between grammatically similar word groups. In this article discusses the use of homonymy between two grammatically similar nouns and adjectives using statistical methods, namely Frequency and Bayesian methods. If bigrams and trigrams are used in the Bayesian method, the characteristics of word groups are classified in the frequentist method, and the parameters that can distinguish them are determined. The identified parameters are converted into numbers as a result of observations and probabilistic decisions are made.

1 Introduction

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.

The phenomenon of homonymy is one of the important elements of the semantic analyzer. 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 method;
- Method based on statistical data;
- A method based on Machine learning.

These methods are used differently in different languages. For example, in Russian linguistics there are many studies devoted to the study of homonyms. The phenomenon of homonymy A.A. Porokhin [2013], D.N. Gomon [2004], D.A. Mikhailovna [2015], P. Boris

* Corresponding author: rector_tsuull@navoiy-uni.uz

Kobritsov, Olga N. Lashevskaja, Olga Ju. Shemanaeva [2011], A.I. Bolshakova [2003], B. A. Bobnev (2010), S.V. Rysakov, E.S. Klyshinsky [2015], A.V. It was specially studied in the works of Gashkov [2012] and others. Statistical methods were used to distinguish homonyms in Russian.

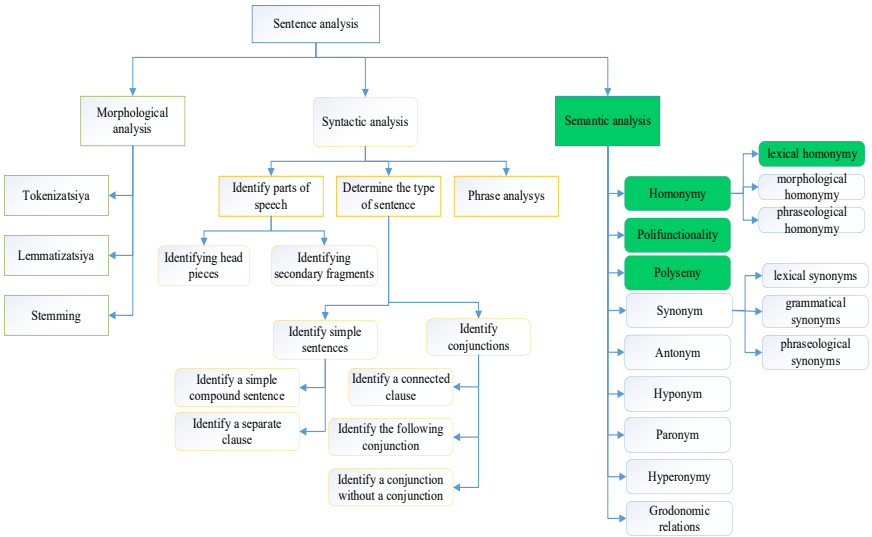

Fig. 1. Linguistic analyzer

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.

2 Experimental part

The phenomenon of homonymy, in turn, is studied by dividing it into such groups as lexical, morphological and phraseological homonymy. This article provides information on the methods used to eliminate lexical homonymy. 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. It is recommended to use rule-based and statistical methods to distinguish homonyms between different word groups. At this point, the question arises: "When is a rule-based method preferable, and when is a statistical method convenient?" In order to use statistical methods, it is necessary to have a natural language corpus with a large aggregate database and all existing texts in it to be tagged. We used the rule-based method to determine homonymy within grammatically dissimilar word groups [2,3,4]. There are also such groups that form homonymy within different word groups that form homonymy within grammatically similar word groups.

The Uzbek language also has its own national corpus, and the database of this corpus contains billions of texts. Therefore, statistical methods can be used. For this, the issue of tagging corpus data, i.e. language modeling [1], is of course cross-cutting.

In this article, we will try to explain the process of differentiation using statistical methods in differentiating homonyms within the word groups presented in Figure 2.

Among the homonyms, there are words that belong to different word families and are united by the same affixes. But after the suffix is added, the constituents of this homonym may belong to different word groups, or the compounds may belong to different hyponyms. In such situations, the Trigram Hidden Markov Model (HMM) is used to distinguish homonyms. To use the trigram HMM, the tags of the words in the sentence must also be determined. Trigram HMM, V is a finite set of possible words and K is a finite set of possible labels of these words, with the following parameters:

.

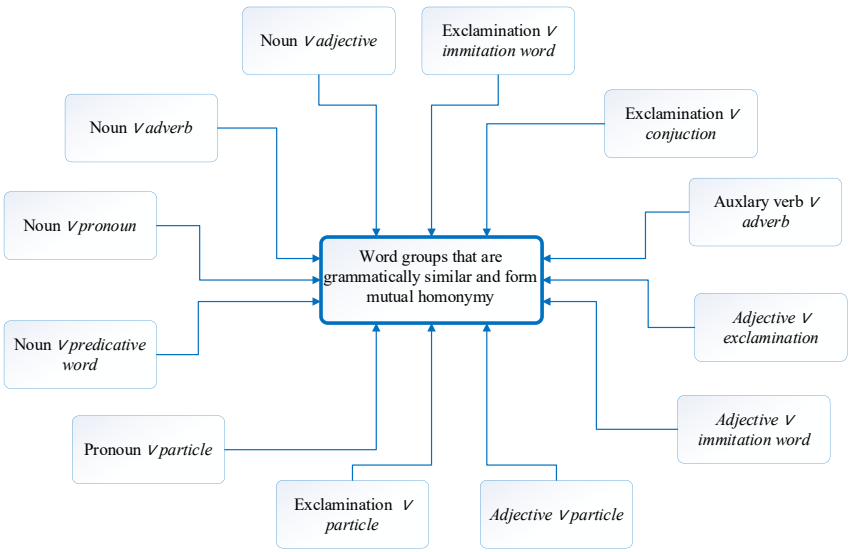

Fig. 2. Word groups that are grammatically similar and form mutual homonymy

– The parameter $q(s|u,v)$ is for every trigram $s|u,v\ s \in K \cup \{STOP\}$ } [1]and $u,v \in V \cup \{*\}$. $q(s|u,v)$ is the probability that the tag s is encountered after the bigram of tags (u,v), * denotes the beginning of the sentence.

– The parameter $e(x|s)$, each $\in V,\ s \in K$. The value $e(x|s)$ determines the probability that word x is paired with word string s.

$S\ (x_1,\dots,x_n,y_1,\dots,y_{n+1}) is$ a collection of pairs of word sequences and label sequences, here, here $n \geq 0,\ x_i \in V,\ i = 1 \dots n,\ y_i \in K\ i = 1 \dots n\ va\ y_{n+1} = STOP$. We have for each $(x_1,\dots,x_n,y_1,\dots,y_{n+1}) \in S$ the following, we need to determine the probability.

$$p(x_1,\dots,x_n,y_1,\dots,y_{n+1}) = \prod_{i=1}^{n} q(y_i|y_{i-2},y_{i-1}) \prod_{i=1}^{n} e(x_i|y_i) \quad (1)$$

Here $y_0 = y_{-1} = *$.

For example, we have n=5, $x_1 \dots x_5$ equal to " *Buvijonning alomat gaplari bor a?*" given a sentence, in order to apply joint probability for this sentence, it is divided into stem and suffixes. When tagging a given sentence, two different values N Adj N STOP and N N N STOP can be formed. The joint probability formula for each set of labels $y_1 \dots y_6$ is calculated as follows. Conditional probability for sequence N Adj N STOP

$$p(x_1, \dots, x_n, y_1, \dots, y_{n+1}) = q(N|*,*) \times q(Adj|*,N) \times q(N|N,Adj) \times q(STOP|Adj,N) \times e(Buvi|N) \times e(alomat|N) \times e(gap|N)$$

This model is a noise-channel model. We use the second-order Markov model (trigram model) to calculate the value of.

$$q(N|*,*) \times q(Adj|*,N) \times q(N|N,Adj) \times q(STOP|Adj,N)$$
$$(Buvi|N) \times e(alomat|N) \times e(gap|N) -$$
$$p(Buvijonning \ \boldsymbol{alomat} \ gaplari \ bor \ a|\ N \ Adj \ N \ STOP)$$

denotes the conditional probability, where p(x|y) is the conditional probability of the x's obtained from the sentence "Grandmother has sign sentences a" and N Adj denotes the conditional probability of the y's obtained from N STOP tags. To calculate this probability, we need some parameters. Now we evaluate these parameters. A sample set $X_1 \dots X_n$ is given for the analyzer. For each sentence, a sequence of $x_1 \dots x_n$ words and $y_1 \dots y_n$ tags is defined. How do we estimate the parameters of the model given this information? It can be seen that there is a simple and very intuitive answer to this question.

c(*u,v,s*)– determines the number of sequences of word groups *u,v,s* in the given data, for example *c(N,Adj,V)* - the homonym of the *alomat* in the given sentence is the adjective word group indicates the number of occurrences of words consisting of tags N and V before and after this word. Similarly, c(u, v) indicates how many times (u,v) meets the bigram characters. And c(s) determines how many times s has been seen in the given data corpus. And finally, c(s↝x) is the number of occurrences of the word x belonging to the s word group in the corpus: for example, c(Adj ↝alomat) is the number of occurrences of the word work in the corpus in the form of the Adj (adjective) tag.

Taking these comments into account, the maximum likelihood is given as follows

$$q(s|u,v) = \frac{c(u,v,s)}{c(u,v)} \quad (2)$$

and

$$e(x|s) = \frac{c(s \rightsquigarrow x)}{c(s)} \quad (3)$$

For example, for our example, this probability is calculated as follows

$$q(N|N,Adj) = \frac{c(N,Adj,N)}{c(N,Adj)}$$

and

$$e(alomat|Adj) = \frac{c(Adj \rightsquigarrow alomat)}{c(Adj)}$$

Thus, to estimate the parameters of the model, it is enough to count numbers from the language corpus with a tagged database and calculate the maximum likelihood using formulas. We perform the above calculations for the sentence *"Buvijonning **alomat** gaplari bor a?"*

3 Results and discussions

70,358 pieces of information were found when searching for the word symptom in the database of the national corpus of the Uzbek language. When the first 500 were analyzed and tagged, the following results were obtained

$$q(N|N,Adj) = \frac{c(N,Adj,V)}{c(N,Adj)} = \frac{228}{351} = 0.65$$
$$e(alomat|Adj) = \frac{c(Adj \rightsquigarrow alomat)}{c(Adj)} = \frac{382}{553} = 0.69$$

Conditional probability value of the sign word and its compounds in the given sentence $p(x_1,\dots,x_n,y_1,\dots,y_{n+1})= q(N|N,Adj) \times e(alomat|Adj) = 0.65 * 0.69 = 0.44$

Similarly, when the conditional probability is calculated for the sequence of tags N N N STOP $p(x_1,\dots,x_n,y_1,\dots,y_{n+1})= q(N|N,N) \times e(alomat|N)$ *=0.56* was found to be equal to From the calculated results, *"Buvijonning alomat gaplari bor a?"*the word symptom in the sentence is a homonymous word of the noun group and can be considered to mean " Odatdagidan o'zgacha tushunib bo'lmaydigan ". This method is called Bayes method. Another statistical method is the frequency method, which requires the classification of word groups. Let's consider the process of determining the above-mentioned *alomat* word using the frequency method.

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

1. Vocabulary;
2. Stem and lemma;
3. Just a lemma
4. Only the stem

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

Bu bemordagi alomatlar covid-19 kasalligini eslatyapti.

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

We use the data of the national corpus of the Uzbek language to classify the word symptom according to the above parameters. A total of *6823* pieces of information were found when searching for the word sign 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.

Based on the obtained statistical data, we also divide the homonym in the given sentence into stems and suffixes

So, based on the above statistics, the following decision is made for the word *alomat* in this sentence.

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, homonymy 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.

4 Conclusion

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:

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

Here, $P(tag|previous\ n\ tags)$ - 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.

References

1. M. Collins, Tagging with Hidden Markov Models, http://www.cs.columbia.edu/~mcollins/courses/nlp2011/notes/hmms.pdf (Last accessed 21.05.2023)
2. 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 (2011)
3. G. I. Kustova, O. N. Lyashevskaya, Ye. V. Paducheva, Ye. V. Raxilina, Semanticheskaya razmetka leksiki v Nasionalnom korpuse russkogo yazika: prinsipi, problemi, perspektivi, Nasionalniy korpus russkogo yazika: 2003-2005. Rezultati i perspektivi, Moscow, Indrik, 155-174 (2005)
4. B. P. Kobrisov, G. I. Kustova, O. N. Lyashevskaya, O. Yu. Shemanaeva, Ye. V. Raxilina, Mnogoznachnost kak prikladnaya problema: semanticheskaya razmetka v Nasionalnom korpuse russkogo yazika, Kompyuternaya lingvistika i intellektualnie texnologii: Trudi mejdunarodnoy konferensii «Dialog-2006», 445-450 (2006)
5. V. V. Kukanova, Prinsipi semanticheskoy razmetki nasionalnogo korpusa kalmiskogo yazika http://kalmcorpora.ru/sites/default/files/kukanova-25.pdf (Last accessed 24.05.2023)
6. A. A. Kretov, Analiz semanticheskix pomet v NKRYa, http://ruscorpora.ru/sbornik2008/11.pdf (Last accessed 27.05.2023)
7. A. Y. Anikin, Opit semanticheskogo analiza praslavyanskoy omonimii na indoevropeyskom fone/ tema dissertasii i avtoreferata po VAK RF 10.02.03, kandidat filologicheskix nauk, Moscow (1983)
8. Y. E. Yermolaeva, Semanticheskoe varirovanie dialektnogo slova v russkix govorax Bashkirii: V svyazi s problemoy razgranicheniya polisemii i omonimii, tema dissertasii i avtoreferata po VAK RF 10.02.01, kandidat filologicheskix nauk, Ermolaeva, Yuliya Aleksandrovna (2000)