UDC:681.51

MATHEMATICAL MODELS THAT DISTINGUISH HOMONYMY IN THE FRAMEWORK OF A WORD SERIES

Axmedova Xolisxon Ilxomovna
PhD student,
Tashkent State University of Uzbek language and literature,
xolisa9029@mail.ru

Annotatsiya. Mazkur maqolada yaratilajak O‘zbek tili semantic analizatorining muhim vazifalaridan biri bo‘lgan bir so‘z turkumi doirasidagi omonim so‘zlarni ma’nolarini farqlovchi omillar, matematik modellar haqida so‘z yuritiladi. Ushbu muammoning yechishda ehtimollar nazariyasiga bevosita murojaat qilamiz. Semantik analiz uchun berilgan matndagi omonim so‘z bilan brikib kelgan so’zlarni aniqlash, aynan shu so‘z bilan brikish ehtimolligini hisoblash masalalari to‘g‘risida ushbu maqolada so‘z yuritiladi.

Kalit so‘zlar. Matematik model, shartli ehtimollik, lingvistik model, omonim, ketma-ketlik, teglar ketma-ketligi, Markov modellari, Trigramm Yashirin Markov modeli.

Аннотации: В этой статье мы поговорим о факторах, которые различают значения омонимических слов внутри категории слов, математических моделях, которые являются одной из важных функций создаваемого семантического анализатора узбекского языка. Решая эту проблему, мы обращаемся непосредственно к теории вероятностей. В этой статье мы поговорим о вопросах определения слов, которые приходят на ум при слове-омониме в тексте, предоставленном для семантического анализа, о вычислении вероятности этого с этим словом.

Ключевые слова: Математическая модель, условная вероятность, лингвистическая модель, омоним, последовательность, последовательность тегов, Марковские модели, триграмма, Скрытая Марковская модель.

Abstract: In this article we will talk about the factors that distinguish the meanings of homonym words within a category of words, mathematical models, which are one of the important functions of the Uzbek language semantic analyzer to be created. In solving this problem, we turn directly to the theory of probability. In this article we will talk about the questions of determining the words that come to mind with the word homonym in the text given for semantic analysis, the calculation of the probability of this with this word.

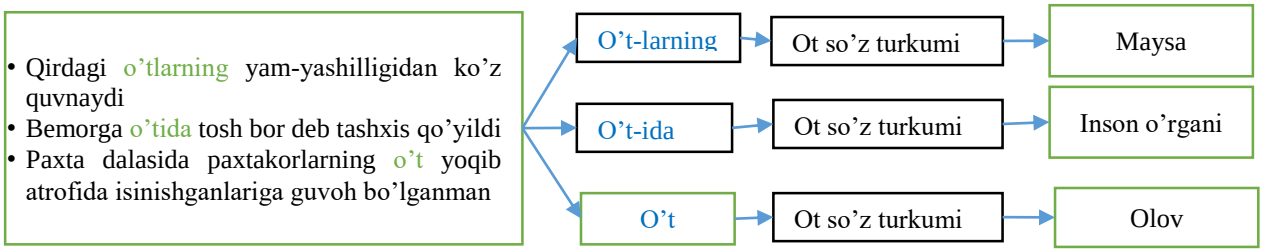
Key words: Mathematical model, conditional probability, linguistic model, homonym, sequence, tag sequence, Markov models, Trigram Hidden Markov model.

Introduction. Linguistic modeling is important in computer linguistics. On the basis of the created model, the software is created and the issue of the issuance of the language unit in the body finds a solution. One of such important tasks of the natural language processing process is the creation of a system that distinguishes the meanings of homonyms. In order to carry out this task through the system, of course, it will be necessary to use linguistic models, mathematical models based on them, algorithms

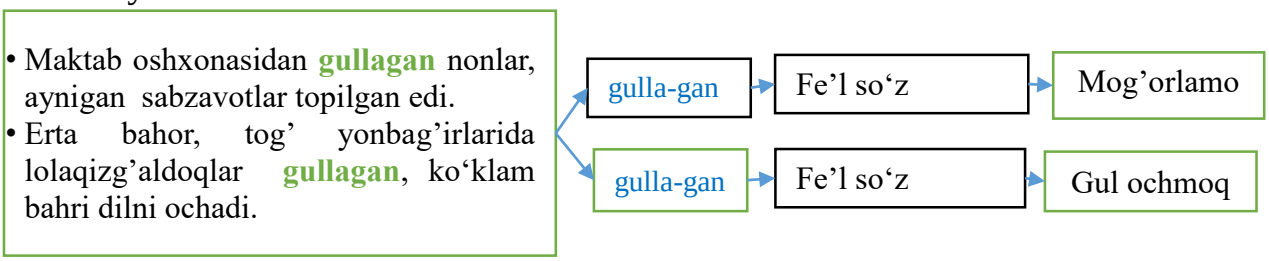
created on mathematical models, database management systems, the programming language for creating the application interface, associating it with the necessary data, and its capabilities. The creation of this system increases the level of excellence of the National Uzbek language Corps.

Literature review. This problem in various foreign language Corps has its own solution. In the English National Corps, homonym was used as a way of touching sentences in differentiating words [1]. This method is much more convenient for English grammar, since there are practically no syntactic and lexical form-building suffixes to English. This is evident in the research of the English scientist Michail Collins. Russian researchers Ladanova E.O. His research has recommended addressing the theory of artificial intelligence in the semantic analysis of mantles. Omonim in the Uzbek language on the words Z. Holmonova, O. Khalyarov, Sh. Gulyamova number of linguists, such as, was in search. In order to distinguish the formative words in the scientific dissertation of Sh. Gulyamova, they were cited linguistic models, which emphasized and distinguished the greater efficiency of determining the circle of categories, O.Kholyorov analyzed until the homonyms between the suffixes. Z. And Kholmonova studied the Explanatory Dictionary of homonym words.

Research Methodology. Having studied the works done in the framework of the foreign and Uzbek language, we propose to divide them into two types in the spiritual differentiation of homonyms in the Uzbek language: different categories of words and homonyms in the same category of words'. And in this article we will consider the process of distinguishing homonyms within the same category of words (noun only, adjective, verb,...). For example:



The noun *o't*, which is presented in these sentences, is a homonym word in the context of the noun constellation, which in each sentence means different meanings. Below are the homonyms in the context of the word series.



It is necessary to model the processes of distinguishing the meanings of homonyms in the context of adjectives, pronouns and other word categories, such as the above examples. And for this, if we look at the linguistic and mathematical methods used in the process of foreign experiments and determination of homonyms in them. Markov models, generative, Noisy channel model, relying on Viterbi algorithms in distinguishing homonymy in English [2-3]. These models can also be used in the Uzbek

language. And for this, it will be necessary to separate each word into its core and suffixes and touch them.

In the sentence given to us for semantic analysis, we divide words into words by the designation x_1, x_2,..., x_n, and these words into cores and suffixes, denoting the tags of the separated cores, that is, within what category of words they are, we define y_1, y_2,..., y_n.

The problem of replacing the x_1, x_2, ..., x_n sequence of sentences into the y_1, y_2, ..., y_n sequence of labels is referred to as the problem of marking or touching the sequence.

Let's assume we have $(x^{(i)},y^{(i)})$ there is a set of examples, (in here, $(x^{(i)}, y^{(i)}), i = 1..m$, each $x^{(i)}$ $Collection$ $x_1^{(i)}, x_2^{(i)}, \dots, x_{n_i}^{(i)}$ consists of sentences and each $y^{(i)}$ AND $y_1^{(i)}, y_2^{(i)}, \dots, y_{n_i}^{(i)}$ a set consisting of a sequence of labels (n_i –i – sample length). Thus $x_j^{(i)}$ while i-in the example j- word, $y_j^{(i)}$ while I is the tag of the J - word in the example.

As X $x_1, x_2, \dots, x_n$ as a set of sequences and Y $y_1, y_2, \dots, y_n$ a set of labels sequences is determined. At the moment our task is to put the sequence of sentences corresponding to the sequence of labels $f: X \to Y$ it consists in learning the function. Sentences in each x source given in machine translation (Chinese e.g.) [3] and each "label" (sign) is perceived as a sentence in the language. Our task is given all the X's and their labels consist of y characters all $(x^{(i)}, y^{(i)})$, *(i=1..n)* it consists in studying the function that determines the.

$f(x)$- Another way to determine the function is to look at the conditional model Method. In this, we define a model that determines the conditional probability of any (x, y) pair.

$$p(y|x)$$

And the model parameters are evaluated using the texts contained in the body. Sample of the result sample from the new x included

$$f(x) = arg \max_{y \in Y} p(y|x) \qquad (1)$$

Y label from this model, the greatest value of the characters. $p(y|x)$ f(x) function is optimal if our model is close to the actual conditional distribution of the given samples. Machine language and natural the fastest used and most alternative model in processing is this generator model. In most cases, we divide the probability p(x, y) by the following.

$$p(x, y) = p(y)p(x|y) \qquad (2)$$

and $p(y)$ and $p(x|y)$ the probabilities for the models will be separate. The components of these two models are interpreted as follows:

$p(y)$ - y Initial probability of distribution of labels.

$p(x|y) - y$ Given that the main tag is, the probability of forming a given x. Joint probabilities p(x) and p(x|y) separation models for conditions Noisy-channel (noisy channel) is called models. Intuitively, we see that the x given as an example is created in 2 different stages:

First, the probability that each y tag is selected p(y);

Second, given X samples p(x|y) the fact that it is formed from distribution.

p(x|y) as a model, if we take the Y tag and assume that the result is X, then this model does not come to hand. Our task, however, is to form y labels with the acceptance of x value.

In conclusion:

- Our task is given x $y = f(x)$ learning the function of a replacement for labels. Examples given to us and values consisting of their labels ($x^{(i)}, y^{(i)}$), *(i=1..n);*
- The Noisy channel (noisy channel) approach is based mainly on using given x examples *p(x)* and *p(x|y)* we make models. And these models are joint models as follows

$$p(x, y) = p(y)p(x|y)$$

- We determine the given x value by using the formula y

$$f(x) = arg\max_{y} p(y|x) \qquad (3)$$

The process of identifying the F(x) tags of those entered in X is also called Decoding. It is considered to calculate the greatest values of this joint probability in distinguishing the meanings of homonymy within the framework of a given series of words from the above. To do this, the Trigram is used from The Hidden Markov model (HMM). The Trigram consists of a finite set of possible words HMM, ϑ and K is a finite set of possible tags of these words, and the following parameters:

- $q(s|u, v)$ Parameter each $s|u, v$ for Trigram $s \in K \cup \{STOP\}$ and $u, v \in \vartheta \cup \{*\}$. $q(s|u, v)$ probability (u, v) after the bigram of the tags, the S tag represents the probability of occurrence, *- denotes the beginning of the sentence.
- $e(x|s)$ Parameters, for each $x \in \vartheta, s \in K$. $e(x|s)$, *x* it determines the probability that the observation will be paired with the S condition.
- S $(x_1, \dots, x_n, y_1, \dots, y_{n+1})$ a collection of pairs of words sequence and sequence of labels, here $n \geq 0, x_i \in \vartheta, i = 1 \dots n, y_i \in K\ i = 1 \dots n$ va $y_{n+1} = STOP$
  We each $(x_1, \dots, x_n, y_1, \dots, y_{n+1}) \in S$ for the following

$$p(x_1, \dots, x_n, y_1, \dots, y_{n+1}) = \prod_{i=1}^{n} q(y_i|y_{i-2}, y_{i-1}) \prod_{i=1}^{n} e(x_i|y_i) \qquad (4)$$

We need to determine the probability. Here $y_0 = y_{-1} = *$.

$$p(x_1, \dots, x_n, y_1, \dots, y_{n+1})$$
$$= q(N|*,*) \times q(N|*, N) \times q(Adj|N, N) \times q(N|N, Adj)$$
$$\times q(V|Adj, N) \times q(STOP|N, V) \times e(Qir|N) \times e(yam - yashil|Adj)$$
$$\times e(ko'z|N) \times e(quvnamoq|V)$$

This model noise-channel

$$q(N|*,*) \times q(N|*, N) \times q(Adj|N, N) \times q(N|N, Adj) \times q(V|Adj, N)$$
$$\times q(STOP|N, V)$$

When calculating the value of *n*, we use the second-order Markov model(trigram model) to determine the probability of the sequence of *N N Adj N V STOP* labels.
$e(Qir|N) \times e(yam - yashil|Adj) \times e(ko'z|N) \times e(quvnaydi|V)$- - $p(Qirdagi\ o'tlarning\ yam - yashilligidan\ ko'z\ quvnaydi|N\ N\ Adj\ N\ V\ STOP)$ denote conditional probability,

In here ***p(x|y)*** "$Qirdagi\ o'tlarning\ yam - yashilligidan\ ko'z\ quvnaydi$" give information *x* and $N\ N\ Adj\ N\ V\ STOP$ indicates the conditional probability of y

from the labels. And for us to calculate this probability, we will need some parameters. Now we will evaluate these parameters. The data given for the analyzer is a set of samples, each sample contains a sequence of $x_1 \dots x_n$ sentences and $y_1 \dots y_n$ tags respectively.

Analysis and results. How do we evaluate the model parameters taking into account the above information? We see that there is a simple and very intuitive answer to this question. c(u, v , s) in the given data, u, v, s determine the number of cases in the sequence, for example, c (N, N, Adj) -when the noun homonym in the given sentence came up in the function of the noun constellation, *the N, N,* of the words that come before and after this word indicates the number of 3 sequences of labels. And so c (he, v) means how many times (he, v) bigram meets the signs. While c(s) determines how many times s has been seen in a given data body.

The maximum-probability, taking into account these comments, is given as follows

$$q(s|u,v) = \frac{c(u,v,s)}{c(u,v)} \qquad (5)$$

and

$$e(x|s) = \frac{c(s \rightsquigarrow x)}{c(s)} \qquad (6)$$

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

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

and

$$e(o't|N) = \frac{c(N \rightsquigarrow o't)}{c(N)}$$

Thus, to evaluate the model parameters, it is sufficient to count the numbers from the Language Body and calculate the maximum probability by formulas. And below we bring a program that will find of the word asked from the given source in the python programming language.

```
    def search_string_in_file( string_to_search):
line_number = 0
list_of_results = []
data_of_line=[]
word_num=0
x=0
with open(sources.txt', 'r', encoding='utf8') as read_obj:
   for line in read_obj:
      line_number += 1
      if string_to_search in line:
         list_of_results.append((line_number, line.rstrip()))
         for j in line.split():
            word_num+=1
            data_of_line.append(j)
            if string_to_search[1:]  in j:
```

```
            x=data_of_line.index(j)
            if(len(data_of_line)==x):
             print(data_of_line[x-1],' ',data_of_line[x])
            else:
             print(data_of_line[x-1],' ',data_of_line[x],' ', data_of_line[x+1])
        data_of_line.clear()
    return list_of_results
```

Conclusion. The above program optionally extracts the required word from the quoted source, as well as a list of words that come before and after this word. Signing up is formed as a result of the program$c(N, N, Adj)$, $c(N, Adj)$, $c(N \rightsquigarrow o't), c(N)$ values are calculated and it is determined that their probability is the greatest value from within. With the greatest probability, the explanation of the word homonym in the database is transferred to the user interface. In this way, the meanings of homonyms in the framework of one word series differ.

References:

[1]. Gulyamova Sh. Axmedova X. O‘zbek tili semantik analizatori uchun omonim so‘zlar ma’lumotlar bazasini shakllantirish masalasi xususida // So‘z san’ati xalqaro jurnali: http://dx.doi.org/10.26739/2181-9297-2021-3-102
[2]. Michael Collins Tagging with Hidden Markov Models //2011.
[3]. Divya Godayal, An introduction to part-of-speech tagging and the Hidden Markov Model // 2018.
[4]. Ladanova E.O. , Yamashkin S.A. Semantic analyzer for the selection of facts from text messages // Mejdunarodniy nauchno-issledovatel'skiy jurnal, international research journal. – Yekaterinburg, 2017.
[5]. Po‘latov A. Kompyuter lingvistikasi. – Toshkent, 2011. – B. 11.
[6]. Rahmatullayev Sh. O‘zbek tili omonimlarining izohli lug‘ati. – Toshkent: O‘qituvchi, 1984. – 108 b.;
[7]. Rahmatullayev Sh.O‘zbek tilining izohli frazeologik lug‘ati. – Toshkent: O‘qituvchi, 1978. –B. 408.
[8]. Russkiy semanticheskiy slovar'. Tolkoviy slovar', sistematizirovanniy po klassam slov i znacheniy / Rossiyskaya akademiya nauk. In-t rus. yaz. im. V. V. Vinogradova; Pod obshey red. N. Yu. Shvedovoy. – M.: Azbukovnik, 1998.