Semantic Differentiation of Uzbek Homonyms Using the Lesk Algorithm

Article · September 2023

CITATIONS
0

READS
13

4 authors:

Botir Elov
Tashkent State University of Uzbek Language and Literature
65 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Axmedova Xolisa
Tashkent State University of Uzbek Language and Literature named after Alisher …
19 PUBLICATIONS **1** CITATION
SEE PROFILE

Primova Mastura
Alisher Navoiy Tashkent state University of Uzbek Language and literature
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Nizomaddin Khudayberganov
Tashkent State university of Uzbek language and literature
7 PUBLICATIONS **1** CITATION
SEE PROFILE

IEEE Turkey Section

UBMK'23

Bildiriler Kitabı
Proceedings

8. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

8th International Conference on Computer Science and Engineering

13-14-15 Eylül (September) 2023 Burdur - Türkiye

Semantic Differentiation of Uzbek Homonyms Using the Lesk Algorithm

Elov Botir Boltayevich
Dept. of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i,
Tashkent, Uzbekistan
elov@navoiy-uni.uz

Axmedova Xolisxon Ilxomovna
Dept. of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i,
Tashkent, Uzbekistan
xolisa9029@mail.ru

Primova Mastura Hakim qizi
Dept. of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i
Tashkent, Uzbekistan
primovamastura@navoiy-uni.uz

Khudayberganov Nizomaddin Uktambay o'g'li
Dept. of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i
Tashkent, Uzbekistan
nizomaddin@navoiy-uni.uz

***Abstract*–The development of a semantic analyzer of natural language is considered one of the factors that develop the language. Homonymy is one of the main elements of semantic analysis. Different methods can be used for semantic analysis of homonyms. Homonyms can also be determined using Lesk's algorithm. Lesk's algorithm is based on WordNet of natural language. The weight of the compounds of the homonymous word in the sentence entered through WordNet is determined. The meaning of the word homonym was determined according to the compounds with high weight.**

Keywords – semantic analyzer, homonymy, rule-based method, statistical method, Lesk's algorithm, word weight, WordNet

I. INTRODUCTION.

The problem of the 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. All new systems for recognizing text, speech, and paralinguistic tools are being developed. Text processing is one of the oldest and most important research in this field. The first studies on automatic text processing belong to the 50s of the XX century. Automatic text processing is divided into several stages, one of which is morphological classification. At this stage, morphological descriptions (gender, number, case, declension, type, etc.) and the initial form of the word called lemma are defined for each word. Morphological classification is complicated by the phenomenon of homonymy.

Although homonymy detection methods based on the use of probabilistic models for texts in some inflected languages are very common, they provide very high accuracy. It has been proven that the Hidden Markov model works better for identifying homonyms in Russian texts.

Semantic search is performed through semantic analysis. The better it is designed, the more effective the search will be. Implementation of semantic analysis directly depends on linguistic resources. Lexical resources include dictionaries, thesauruses, and ontologies. Semantic analysis also has elements, which require a separate study. This article talks about the solution to the problem of identifying homonymy. Homonymy is one of the important elements of semantic analysis. Homonymy detection is interpreted differently in different natural languages. In world computer linguistics, 3 methods are mainly used in the semantic analysis of sentences:

A rule-based method is the detection of homonymy based on predetermined language rules based on the grammatical properties of natural language.

A method based on statistical data can also be called a decision-making method based on the data of the language corpus. That is, statistics are obtained based on the observations made among the data in the language corpus. A new homonym is evaluated based on the received statistical data. The problem of identifying homonymy using statistical methods finds its solution in the process of solving the problem of POS (Past of Speech) tagging of sentences. POS tagging is the process of associating each word in the newly entered text with the appropriate POS tags, such as nouns, verbs, adjectives, etc. This task is one of the meaning recognition tasks in NLP (Natural Language Process). This is because many words in the language can have several different meanings.

Therefore, the use of statistical methods in identifying homonyms between different parts of speech gives effective results.

A method based on machine learning is a method that not only uses statistical data but also directly refers to neural networks. The approach based on machine learning, in turn, is divided into Supervised and Unsupervised algorithms. Good results can also be obtained by using this approach in identifying homonymy.

The meaning of the word can be determined using these methods. Determining the meaning of words in natural language is a complex and important task. Defining the meaning of the word allows for increasing the level of accuracy of the machine translation and creates an automatic annotation for the text, works, scientific articles, and dissertations. One of the most important tasks in defining the meaning of a word is distinguishing the homonym words in a semantic way. Finding out the meaning of the homonym word that is in a sentence.

II. MATERIALS AND METHODS

Having deeply studied foreign experiences, we use rule-based, stochastic, machine learning, and neural network methods to distinguish Uzbek homonyms. When distinguishing homonyms in the Uzbek language, we divided them into groups such as homonyms within one part of speech, two parts of speech, three parts of speech, and four parts of speech according to their occurrence within parts of speech. We used a rule-based method to determine homonymy within grammatically dissimilar word groups. We have mentioned this in scientific articles cited [1],[2], [3].

The problem of identifying homonymy between different word groups is solved in the process of POS (Part of speech tagging) tagging of sentences. Many scientific articles can be found on the use of the Hidden Markov model in the identification of word groups. In the process of using the Markov model, it can be seen that the results will be more accurate using the Viterbi algorithm.

In different natural languages, there are also homonyms that occur within the same word family.

TABLE I. HOMONYMS WITHIN A POS

Words	POS	Senses
O't	Verb	O'tmoq fe'li (cross)
	Noun	Maysa, o't-o'lan (plants)
	Noun	Inson o'rgani (gall bladder)
	Noun	Olov (fire)
Oz	Adverb	Kam, miqdori nisbatan ko'p bo'lmagan (few)
	Verb	Oriqlamoq, etidan yo'qotmoq(lose weight)
	Verb	Noxush bo'lmoq, kuchsizlanmoq, holsizlanmoq (weaken)
	Verb	Adashmoq, to'g'ri yo'ldan chetga chiqmoq (to lose one's way)
Bo'y	Noun	Uzunlik o'lchovi (height)
	Noun	Hid, is (smell)
…	…	…

As shown in Table 1, there may be words that can form homonyms within one part of speech or even among different word groups. Methods such as Frequentist, Naïve Bayes, and Hidden Markov model can be used to determine the word group of homonyms [4].

However, it is not effective to use these methods in the semantic differentiation of homonyms within the same part of speech. It is important to determine what exactly these words mean in a sentence. Solving this problem helps to increase the accuracy of machine translation, to determine the summary of the sentence, sentence-by-text, and text-by-text summary of larger works. The issue of determining the meaning of homonymy words within a word group is called word sense disambiguation (WSD) in foreign sources.

Word meaning discovery (WSD) techniques are useful for many NLP tasks that require semantic interpretation of input. In addition, such methods help to estimate the frequency of different meanings of words in different corpora, which is important for lexicographic research and language learning resources. Although previous research on the meaning of polysemantic verbs in Russian has yielded some important and interesting results, it has mainly focused on reducing ambiguity or identifying frequent meanings. was aimed, but not at assessing the accuracy of word sense disambiguation (WSD). According to the data, there is no comprehensively evaluated method that performs semi-supervised word meaning recognition for Russian verbs. Different variants of the method can be compared and its limitations analyzed. Lexical-semantic ambiguity is a characteristic feature of any natural language, so word sense disambiguation (WSD) is an important part of many natural language processing tasks. Various WSD algorithms have been discussed by Pradhan et al. in sessions of different versions of SemEval [5] and WSD queries have been substantiated in the scientific works of many scientists such as Ide and Veronis [6], Navigli [7]. *The most modern and promising approaches are those that use already existing resources and do not require human input.* Rule-based approaches use thesaurus. Unsupervised corpus-based approaches typically perform clustering on a corpus without explicit reference to an inventory of meanings. We will consider the process of using the LESK algorithm to determine the meaning of words based on corpus data.

III. MAIN

The Lesk algorithm was introduced by E. Michael Lesk in 1986 and is a classic WSD algorithm. Lesk's algorithm is based on the idea that only words in a certain part of the text have a similar meaning. In the simplified Lesk algorithm, the correct meaning of each word context is found by finding the most similar meaning between the given context and its dictionary meaning. This algorithm was used to determine the meaning of a word in Hindi. It was used in the development of question-answer systems, and sentiment analysis systems [8]. Used by Zouaghi, A., Merhbene, L. and others to determine the meaning of Arabic words. 73% accuracy was achieved during the conducted research [9]. Basuki, S., Kholimi, A. S. et al. used semantic identification of Indonesian homographs. As a result, 78.6% accuracy was achieved when identifying words in one-word group, and 62.5% when identifying homonymous words in two-word groups [10]. Lesk's algorithm is a rule-based method based on WordNet of natural language. The following information is required to determine homonymy using Lesk's algorithm:

Set of senses: A collection of existing meanings of a polysemous word.

A set of contexts: a set of contexts for each meaning of a word.

The process of identifying homonymy using Lesk's algorithm consists of two modules.

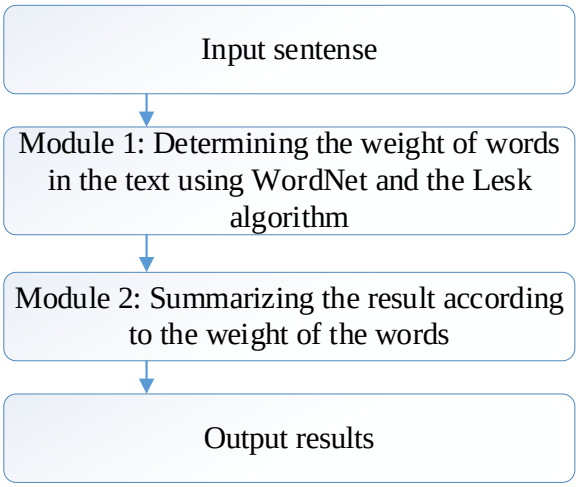

Fig. 1. Modules in the Lesk algorithm

Module 1: The next step is to create a context for each meaning of these words. Corpus data is used for this purpose. Initially, contexts containing homonyms are separated and

they are separated depending on the meaning of the word amonim. Preprocessing is performed on the allocated contexts. (Fig. 2).

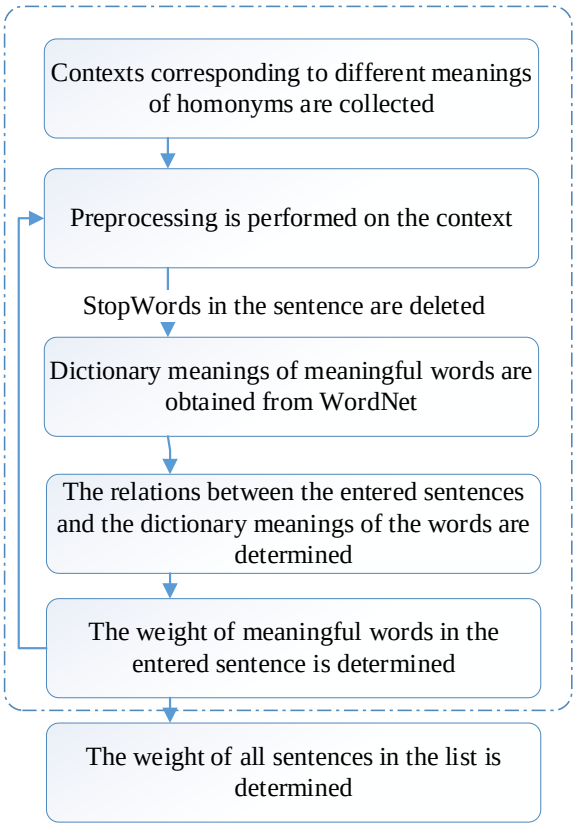

Fig. 2.Tasks of Module 1

Module 1: this algorithm takes O(n3) time for n sentences and performs $O(n^2)$ operations to determine the number of similar words in the sentences.

In Module 2, the results of Module 1 are summarized and summarized.

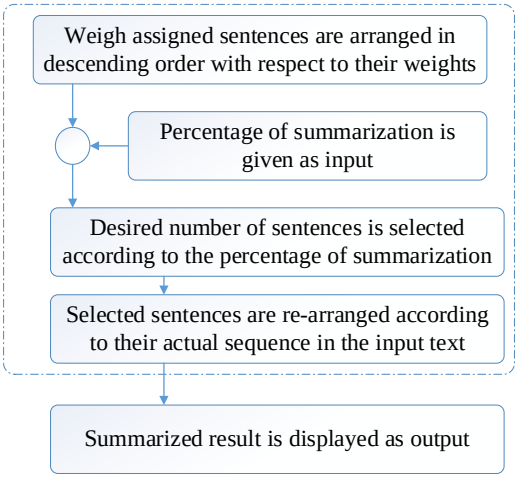

Fig. 3: Tasks of module 2

Let's consider the sequence of semantic differentiation of ambiguous words using the Lesk algorithm using the following sentence. “Bu fe’ling bilan hammani qon qilasan-ku”

The given sentence contains the homonyms "fe’l" and "qon". A set of meanings of these words is included in the database (Table 2).

TABLE II. HOMONYMS IN THE SENTENCE AND THEIR MEANINGS

Words	POS	Senses
Fe’l	Ot	Harakat, xulq-atvor, xarakter
		Grammatik termin
qon	Ot	Organizm tomiridan yurak faoliyati bilan harakatlanuvibi ta’minlovchi qizil rangli suyuqlik
	Fe’l	To’ymoq, qoniqmoq

When determining the weight of words, it is done by counting the number of words for each meaning. The information obtained by the contexts is stored in the database. Above

“Bu fe’ling bilan hammani qon qilasan-ku”

We determine the meaning of homonyms in the sentence using Lesk's algorithm. The following actions are performed on the entered sentence.

Tokenization;
Lemmatization;
Remove StopWords;
POST tagging;
As a result

Fe’l, bilan, qon, qilmoq

The words remain. In the next step, the homonymous word in the entered sentence is determined and its compounds are separated. The weight of the separated compounds is determined from the database. For example, a dataset consisting of “fe’l” conjugations and their weights is presented in Table III.

TABLE III. SEMANTIC COMPOUNDS OF THE ‘FE’L’ AND THEIR WEIGHT

Compounds word	Sense-1	Sense-2
Yomon	10	2
Yaxshi	15	1
Ot	0	25
Qurmoq	35	0
Tor	18	8
Qon	20	12
Qilmoq	14	18

In the same way, a dataset consisting of compounds of the word *“qon”* homonym and their weights is extracted.

TABLE IV SEMANTIC COMPOUNDS OF THE ‘QON’ AND THEIR WEIGHT

Compounds word	Sense-1	Sense-2
Suv	10	35
Tomir	123	0
Kengaymoq	45	5
Fe'l	47	26
Qilmoq	31	11
…	…	…

From the given information, the compounds of the homonymous word in the sentence and their weight are determined. From the determined data, the meaning of the compound with the highest weight is selected.

The weight of the sentence is determined by the weight of each of the compounds present in the sentence. Since this

event is a joint event, the weight of the sentence is equal to the product of the weights.

$$p = \prod_{i=1}^{n} p_i$$

Here, n is the number of words in each sentence, and p_i is the probability that the i- word in the sentence means one meaning. This probability is also known as conditional probability. This conditional probability is calculated for each word in the sentence and the probability of the sentence is determined by generalization

IV. CONCLUSION

Homonymy detection using the Lesk algorithm is also a component of the rule-based method. To use this algorithm, the WordNet system of natural language must be available. More precisely, the Lesk algorithm works based on WordNet. In order to determine homonymy using the Lesk algorithm, homonym words are semantically tagged with the help of the human factor in contexts according to each meaning. The task of this algorithm is to determine and conclude the number of unique words in a semantically tagged context. Lesk's algorithm is a part of the rule-based method, with the help of which it is possible to semantically distinguish not only homonyms from different word groups, but also homonyms from the same word group. Using this algorithm, a large collection of sentences with homonyms is needed. This algorithm can be used with a set of 200-2000 sentences for each homonym.

REFERENCES

[1] E. B. Boltayevich & A. X. Ilxomovna, (2022). Business Process Modeling That Distinguishes Homonymy Within Three Parts of Speechs in The Uzbek Language. In Proceedings - 7th International Conference on Computer Science and Engineering, UBMK 2022 (pp. 278–283). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/UBMK55850.2022.9919453

[2] B.B Elov, X. I. Axmedova, Uchta so‘z turkumi doirasidagi omonimiyani farqlovchi biznes jarayonni modellashtirish// O’zbekiston respublikasi innovatsion rivojlanish vazirligining, Ilm-fan va m innovasion rivojlanish ilmiy jurnal 2022 / 1, 150-162-b.

[3] X.I. Axmedova, Turli so‘z turkumlari orasidagi omonimiyani aniqlovchi matematik modellar// Science and innovation international scientific journal volume 1 issue 7 uif-2022: 8.2 | issn: 2181-3337. https://doi.org/10.5281/zenodo.7238546

[4] B.B. Elov, X.I. Axmedova “So‘Z Ma’nosini Aniqlashda Naive Bayes Algoritmidan Foydalanish”, Илм-Фан Ва Инновацион Ривожланиш ilmiy jurnali, Toshkent, 3/2023,44-54-b.

[5] P. Sameer, L.Edward, D. Dmitriy, and M. Palmer. 2007. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 87–92,

[6] N. Ide and J. Véronis. Word Sense Disambiguation: The State of the Art. Computational Linguistics, 1998, 24(1), 1-41-p.

[7] R. Navigli, Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41.2: 10. 2009.

[8] P. Tripathi, P. Mukherjee, M. Hendre, M. Godse,& B. Chakraborty, (2021). Word Sense Disambiguation in Hindi Language Using Score Based Modified Lesk Algorithm. International Journal of Computing and Digital Systems, 10(1), 939–954. https://doi.org/10.12785/IJCDS/100185

[9] A. Zouaghi, L. Merhbene,& M. Zrigui, (2012). Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation. Artificial Intelligence Review, 38(4), 257–269. https://doi.org/10.1007/s10462-011-9249-3

[10] S. Basuki, A.S. Kholimi, A.E. Minarno, F.D.S. Sumadi, & M.R.A. Effendy,(2019). Word Sense Disambiguation (WSD) for Indonesian homograph word meaning determination by LESK Algorithm Application. In Proceedings of 2019 International Conference on Information and Communication Technology and Systems, ICTS 2019 (pp. 8–15). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICTS.2019.8850957

[11] Elov B.B., Axmedova X.I. Determining homonymy using statistical methods.// “Hisoblash modellari va texnologiyalari (HMT 2022)” O‘zbekiston-Malayziya ikkinchi xalqaro konferensiyasi materiallari-Toshkent, 2022 16-17 sentabr,-106 b.

[12] Uri Roll, Ricardo A. Correia, Oded Berger-Tal// Using machine learning to disentangle homonyms in large text corpora- Conservation Biology 31 October 2017 https://doi.org/10.1111/cobi.13044

[13] J.Y. Park, Shin, H.J.; Lee, J.S. Word Sense Disambiguation Using Clustered Sense Labels. Appl. Sci. 2022, 12, 1857. https://doi.org/ 10.3390/app12041857