ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫНЫҢ
әл-Фараби атындағы Қазақ ұлттық университетінің

Х А Б А Р Л А Р Ы

ИЗВЕСТИЯ

НАЦИОНАЛЬНОЙ АКАДЕМИИ НАУК
РЕСПУБЛИКИ КАЗАХСТАН
Казахский национальный университет
им. аль-Фараби

N E W S

OF THE NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
Al-Farabi
Kazakh National University

SERIES
PHYSICO-MATHEMATICAL

5 (333)

SEPTEMBER – OCTOBER 2020

PUBLISHED SINCE JANUARY 1963

PUBLISHED 6 TIMES A YEAR

ALMATY, NAS RK

«ҚР ҰҒА Хабарлары. Физика-математикалық сериясы».
ISSN 2518-1726 (Online), ISSN 1991-346X (Print)
Меншіктенуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.).
Қазақстан Республикасының Ақпарат және коммуникациялар министрлігінің Ақпарат комитетінде 14.02.2018 ж. берілген **№ 16906-Ж** мерзімдік басылым тіркеуіне қойылу туралы куәлік.

Тақырыптық бағыты: *физика-математика ғылымдары және ақпараттық технологиялар саласындағы басым ғылыми зерттеулерді жариялау.*

Мерзімділігі: жылына 6 рет.
Тиражы: 300 дана.

Редакцияның мекенжайы: 050010, Алматы қ., Шевченко көш., 28; 219, 220 бөл.; тел.: 272-13-19; 272-13-18, http://physics-mathematics.kz/index.php/en/archive

---

Типографияның мекенжайы: «NurNaz GRACE», Алматы қ., Рысқұлов көш., 103.

«Известия НАН РК. Серия физика-математическая».
ISSN 2518-1726 (Online), ISSN 1991-346X (Print)
Собственник: РОО «Национальная академия наук Республики Казахстан» (г. Алматы).
Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и коммуникаций Республики Казахстан **№ 16906-Ж**, выданное 14.02.2018 г.

Тематическая направленность: *публикация приоритетных научных исследований в области физико-математических наук и информационных технологий.*

Периодичность: 6 раз в год.
Тираж: 300 экземпляров.

Адрес редакции: 050010, г. Алматы, ул. Шевченко, 28; ком. 219, 220; тел.: 272-13-19; 272-13-18,
http://physics-mathematics.kz/index.php/en/archive

Адрес типографии: «NurNaz GRACE», г. Алматы, ул. Рыскулова, 103.

News of the National Academy of Sciences of the Republic of Kazakhstan. Physical-mathematical series.
ISSN 2518-1726 (Online), ISSN 1991-346X (Print)
Owner: RPA "National Academy of Sciences of the Republic of Kazakhstan" (Almaty).
The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Communications of the Republic of Kazakhstan **No. 16906-Ж**, issued on 14.02.2018.

Thematic scope: *publication of priority research in the field of physical and mathematical sciences and information technology.*

Periodicity: 6 times a year.
Circulation: 300 copies.

Editorial address: 28, Shevchenko str., of. 219, 220, Almaty, 050010, tel. 272-13-19; 272-13-18,
http://physics-mathematics.kz/index.php/en/archive

Address of printing house: «NurNaz GRACE», 103, Ryskulov str, Almaty.

NEWS
OF THENATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X **https://doi.org/10.32014/2020.2518-1726.84**
Volume 5, Number 333 (2020), 68 – 75

D. Rakhimova[1,2], A.Turganbayeva[1]

[1]Kazakh National University named after al-Farabi, Almaty, Kazakhstan;
[2]Institute of Information and Computational Technologies, Almaty, Kazakhstan.
E-mail: di.diva@mail.ru, turganbaeva.aliya@bk.ru

SEMANTIC ANALYSIS OF THE KAZAKH LANGUAGE BASED ON THE APPROACH OF NEURAL NETWORKS

Abstract. This paper provides an overview of existing modern methods and software approaches for semantic analysis. Based on the research done, it was revealed that, for the semantic analysis of text resources, an approach based on machine learning is most used. This article presents the developed algorithm for the semantic analysis of the text in the Kazakh language. The paper also presents a software solution to this approach implemented in the Python programming language. The vector representation of words was obtained by machine learning based on the corpus, which is 1 million sentences in the Kazakh language. In the software implementation, well-known libraries such as gensim, matplotlib, sklearn, numpy, etc. were used. Based on a set of semantically related pairs of words, an ontology for a specific document is built, which is formed during the operation of a neural network. The paper presents the results of the experiments in the graphical form of a set of words. The novelty of the proposed approach lies in the identification of semantic close words in meaning in texts in the Kazakh language. This work contributes to solving problems in machine translation systems, information retrieval, as well as in analysis and processing systems in the Kazakh language.

Keywords: word2vec, model, vector, word, representation, semantic, analysis, Kazakh, language.

1 Introduction

Computer semantic analysis is closely related to the problem of text understanding by a machine. There are many interpretations of the concept "meaning of the text" and the tasks of understanding it. For example, according to D.A. Pospelov [1], the system understands the text entered into it if from the point of view of a person (or a group of experts) it correctly answers questions related to the information contained in the text. Here we can talk not about simply obtaining facts that are clearly present in the text, but about revealing the hidden meanings that the author introduces. D.A. Pospelov identifies several levels of text comprehension, from the point of view of the complexity of the questions that the intellectual system should be able to answer. Guided by the definition from [1], the meaning of the text can be considered as a description of the knowledge contained in it, in the formal language of knowledge representation, which allows solving a fairly wide range of problems related to text analysis, and the problem of semantic analysis - as a translation of a natural language - expressions into the language of knowledge representation. For example, the language of first-order predicates, semantic networks, frames, as well as ontologies and thesauri can act as a language for representing knowledge of a text in a natural language.

In the 60s and 70s, the main approach to representing the semantics of a language was the component approach, within which the meaning of each word in a natural language had to be represented as a combination of semantic universals. By the mid-1980s, it became clear that a generally accepted set of such universals had never been compiled. Relational semantics has become an alternative to the component approach in semantics. In this approach, the meanings of the words of the language are described by setting connections with the meanings of other words, and the entire conceptual system of the language is represented as a semantic network [2].

Review of methods and software approaches for semantic analysis. Of course, no software can replace the analysis that humans can think of. However, the programs that are currently being developed can reduce the time spent studying large databases. In this regard, the work of the following programs for

solving problems of semantic text analysis is considered. Software offered by various manufacturers, such as “Semantic LLC”, “Tomita-parser (Yandex)”, Semantic Analyst “JHON”, “SummarizeBot API”, “TextAnalyst 2.0”, “Galaktika-ZOOM”, “NLP ISA”, “Natasha” and etc. is used in various subject areas and for different languages [3-10]. A complete overview of existing modern systems of semantic analysis and their description are presented in table 1.

Table 1 – Review of modern software systems for semantic text analysis

System name	Description
“Semantic LLC”	is a program for editing unstructured text. The semiconductor line is graphically oriented, each node is a semantic element, and the walls represent the elements of the elements. Each node attribute has a great value, the set of attributes depends on the element type.
“Tomita-parser (Yandex)”	a program that allows you to extract facts from structured text. Separation of facts is based on context-independent grammar rules. And the program requires a dictionary of keywords. The parser will write its own grammar.
“JHON”	The semantic analyst "*JHON*" receives the meanings of a natural language in Russian and solves the following tasks: lexical analysis, morphological analysis, syntactic analysis, semantic analysis - involving the triad of subject-object relations, creating a semantic network of text, fact of events.
“SummarizeBot API”	The web service offers a RESTful API to handle all text and image processing tasks. It uses over 100 languages including Russian, English, Chinese, Japanese and uses machine learning technology. The current version uses the following parameters: 1) automatically link to text; 2) Selection of keywords and conceptual documents; 3) Analysis of a sample of documents and selection of material objects and attributes; 4) Automatically detect the language of the document; 5) Obtaining unpublished data: the main text of articles, forums, forums, etc.; 6) Image processing: identification and recognition of objects in images.
“TextAnalyst 2.0”	the program was developed by the research and production innovation center MicroSystems as a tool for text analysis. Text links allow you to create a semantic web of comments, expressed in processed text. The request has the ability to semantic search for text fragments, taking into account the semantic links hidden in the text. Allows you to parse text by composing a hierarchical tree / heading topics containing text.
“Galaktika-ZOOM”	an automated information search and analysis system manufactured by the Galaktika Corporation. It is a powerful editing and processing tool that allows you to get the information you need in large quantities. It is offered as a commercial system with consumers in advertising, government, and media. This program allows you to build semantic networks, but its program codes are not shared with the system.
“NLP ISA”	For the text, a tree of analyzed analysis was built, the semantic role and connections were established. Allows you to select serialized syntax and semantic analysis mode. Alternatively, you can also select a mode that has a syntax-semantic mode combination.
“Natasha”	it is a set of rules for getting a Tomita parser for Python and a set of ready-to-execute rules, addresses, terms, sums and other objects.

Scientific works [11-13] describe the basic ideas of information retrieval. Various options for finding text statistics are presented, which include counting the number of occurrences of words in documents and the frequency of word contiguity, and new model architectures for computing continuous vector representations of words from very large datasets. The quality of vector representations of words obtained by various models was studied using a set of syntactic and semantic language problems. In [14], the application of language models of a neural network to the problem of calculating semantic similarity for the Russian language is shown. The tools and bodies used, and the results achieved are described.

The above software products are designed for multi-resource languages such as English, Spanish. Russian, etc. Unfortunately, for the Kazakh language now there is no software implementation in the open access. This is since the Kazakh language differs in its semantic and linguistic properties from others, and also does not have large linguistic resources for conducting applied research.

2 Algorithm for semantic analysis of text in the Kazakh language

During digital technologies, given the constant growth of the volume of digital data, an important role is played by improving the quality of information retrieval using new semantic approaches and methods.

To work with big data, various algorithms and methods are being developed for the machine solution of this problem, since the amount of data does not allow for manual analysis. Any natural language is

complex, unique, and multifaceted in its own way, therefore, extracting data from documents and text resources is a large and time-consuming work that requires preliminary processing.

Based on the research done from the developed models used most for the semantic analysis of text resources, there is an approach based on machine learning. Below will be presented the developed algorithm for semantic analysis of text in the Kazakh language and implementation based on this approach. When developing an algorithm to map certain information to a certain attribute, we opted for a neural network (NN) with a hidden layer (100). Neural network training consists of the following parts:

– Text preprocessing. Text preprocessing consists of three stages: tokenization, removal of stop words, normalization of words.

– Construction of the feature vector. The feature vector is a sign of the characteristic we are interested in. For one descriptor, the features were taken as follows: a window of two words after, five before was taken in the text of the article at the place of occurrence of the element. Moreover, a dictionary is formed for each descriptor, which is responsible for the presence of the specified word in the dictionary. All features of each descriptor are collected into one and a feature vector is constructed.

– Training the neural network. The network is trained by presenting each input dataset and then propagating the error.

At the second stage, the neural network was trained. For text preprocessing, the developed natural language processing modules were used. After applying these modules, we extracted the features of our descriptor. A feature vector was then constructed using the extracted data. The constructed feature vector was compared with certain keywords, determined by the modified TF-IDF method for the Kazakh language.

3 Software solution and algorithm implementation

This is one of the most difficult and demanded tasks facing artificial intelligence is NLP (Natural Language Processing). To solve and implement NLP tasks currently, there are several software systems and libraries, which include the tasks of speech recognition, language formation and information acquisition, etc.

Python is currently one of the most promising programs for solving NLP problems. Libraries written in Python are designed to solve NLP problems and allow you to simulate various languages and processing functions.

There are also many types of libraries, consider the most famous and applicable for word processing tasks:

Spacy, NLTK, CoreNLP, StanfordNER, etc. Table 2 below shows a comparison of the functional capabilities for solving the NLP problem.

Table 2 – Comparison of the capabilities of libraries aimed at solving NLP problems

Function	Spacy	NLTK	CoreNLP
Programming language	Python	Python	Java/Python
Neural network models	+	-	+
Vector of integrated words	+	-	-
Multilingual model	+	+	+
Tokenization	+	+	+
POS tagging	+	+	+
Segmantation	+	+	+
Parsing	+	-	+
Highlighting named objects	+	+	+
Communication between objects	-	-	-

Having studied the technical possibilities for the implementation of the semantic analysis algorithm and training the neural network, the authors will use the Spacy and StanfordNER libraries. The StanfordNER and Spacy libraries allow us to model our own model. It also allows you to make the necessary configurations, depending on the specifics of the (Kazakh) language in question.

It is necessary to define the StanfordCoreNLP settings [15]: token- tokenize; ssplit - distribution of offers; pos - speech definition; lemma - find the original form of each word; ner - highlighting named objects; - regexner - work with regular expressions; parse - semantic analysis of each word; depparse - definition of syntax between words and sentences.

Further, figure 1 shows the developed algorithm for the implementation of semantic analysis taking into account keywords and describes the work of the modules.

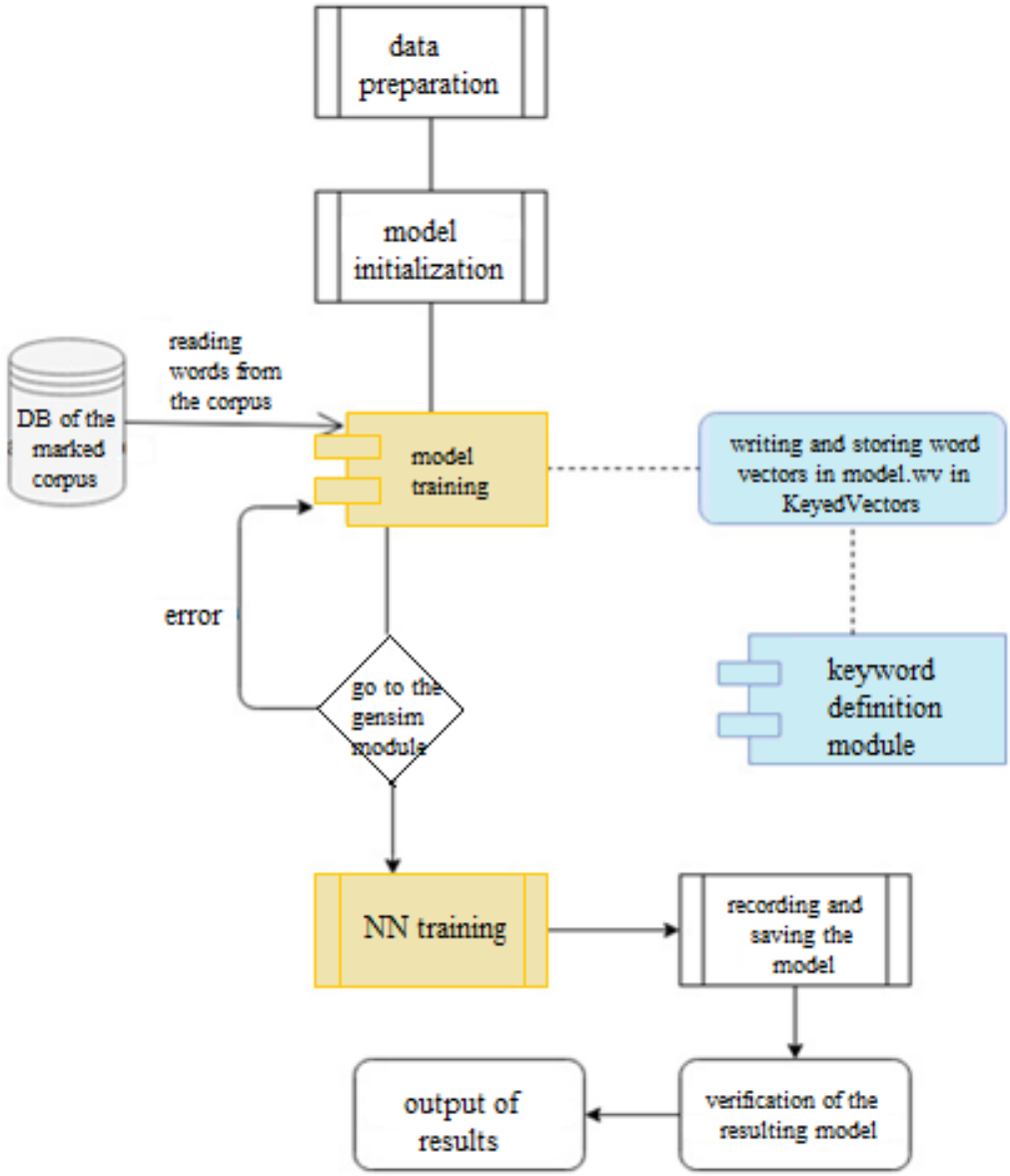

Figure 1 – Algorithm for the implementation of semantic analysis taking into account keywords

The input is text data. To train the model, set the following parameters: The dimension of the feature vectors is 100; The maximum distance between the current and predicted word in a sentence is 5; The minimum education level is 1; The cutoff frequency is 4 words.

```
>>> model = WordVec(sentences, size=100, window=5, min_count=5, workers=4)
```

Record initialized model

```
>>> model.save(fname)
>>> model = WordVec.load(fname) #
```

Now you can train with the resulting model. For training the model, a monolingual Kazakh corpus was prepared, which is in the SQL database. When the text is processed by the model, vectors of words are identified, which are stored in the model.wv module in KeyedVectors. The resulting vectors of words are also compared with keywords (phrases) from the text corpus for the purpose of further use as possible values of semantic attributes of entities. Once the model finishes training, you can go to gensim.models.KeyedVectors in wv:

```
>>> word_vectors = model.wv
>>> del model
```

The gensim.models.phrases module automatically detects a long chain of words. This module allows us to define phrases through learning.

```
>>> bigram_transformer = gensim.models.Phrases(sentences)
>>> model = Word2Vec(bigram_transformer[sentences], size=100, ...)
class gensim.models.wordvec.Corpus(dirname)
class gensim.models.wordvec.LineSentence(source,max_sentence_length=10000, limit=None)
```

After completing the gensim module, you can then start training the neural network.

```
sentences = LineSentence('myfile.txt')
from gensim.models import Word2Vec # define training data
sentences = [['ұл (ul)', 'балалар (balalar)', 'қыздарға (qyzdarga)', 'қарағанда (qaraganda)', 'мықты (myqty)', 'болады (bolady)'],
['Ал (Al)', 'қыз (qyz)', 'балалар (balalar)', 'ұлдарға (uldarga)', 'қарағанда (qaraganda )','нәзік (nazik)'], ['Қыз (Qyz)', 'әлемнің (alemning)', 'көркі (korki)'],
['Гүл (Gul)', 'жердің (zherding)', 'көркі (korki)'],
['Қазақстан (Kazakhstan)', 'республикасы (respublikasy)', 'тәуелсіз (tauelsiz)', 'мемлекет (memleket)']]…
```

As a result of the obtained trained model, it is necessary to check the obtained data. You can also create a graphical interpretation of the results (Figure 2).

The NER Stanford software package was used to train the model. The following is a listing of working with the Stanford NER library and software implementation

```
>>> trainFile = train/dummy-kazakh-corpus.tsv serializeTo = dummy-ner-kazakh-french.ser.gz map = word=0,answer=1
useClassFeature=true useWord=true useNGrams=true noMidNGrams=true maxNGramLeng=6 usePrev=true useNext=true
useSequences=true usePrevSequences=true maxLeft=1 useTypeSeqs=true useTypeSeqs2=true useTypeySequences=true wordShape=chris2useLC useDisjunctive= true
```

Figure 2 - Graphical representation of the vector space of practical results of the semantic analysis of the text in the Kazakh language

```
>>> cd stanford-ner-tagger/
java -cp "stanford-ner.jar:lib/*" -mx4g edu.stanford.nlp.ie.crf.CRFClassifier -prop train/prop.txt
```

```
# coding: utf-8

import nltk
from nltk.tag.stanford import StanfordNERTagger

# Optional
import os
java_path = "C:\Program Files (x86)\Java\jdk1.8.0_201"
os.environ['JAVA_HOME'] = java_path

sentence = u"Қазақстанда алма өседі. Алматы қаласында ҚазНУ жоғары оқу орны орналасқан"

jar = './stanford-ner-tagger/stanford-ner.jar'
model = './stanford-ner-tagger/my-ner-model-french.ser.g'

ner_tagger = StanfordNERTagger(model, jar, encoding='utf8')

words = nltk.word_tokenize(sentence)
print(ner_tagger.tag(words))
```

Figure 3 – An example of input data of text for the program .NER

The problem was successfully solved by using a morphological parser for marking up parts of speech in texts with the subsequent application of the machine learning method of semantically related keywords (phrases). A trained neural network with a hidden layer is applied to the set of these phrases in order to assign a specific phrase to a specific attribute of the entity described in the text. Thus, based on a set of semantically related pairs of words, an ontology is built for a specific document, which is formed during the operation of a neural network.

Conclusion

To solve the problem, semantic analysis in the Kazakh language is based on machine learning. The program is implemented in the python programming language, using the libraries gensim, matplotlib, sklearn, numpy, etc. A set of vectors of words in the Kazakh language was obtained, which was trained on the corpus, which is 1 million sentences. The corpus is fed to the program input in a normalized form. Further, to improve the result, the corpus will be supplemented with proposals on various topics.

Acknowledgments

The study was supported by the Ministry of Education and Science of the Republic of Kazakhstan within the framework of the scientific project AP 05132950 "Development of an information-analytical data retrieval system in the Kazakh language".

Д.Р. Рахимова[1,2], Ә.О. Тұрғанбаева[1]

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2]Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан

НЕЙРОНДЫҚ ЖЕЛІЛЕРГЕ НЕГІЗДЕЛГЕН ҚАЗАҚ ТІЛІНІҢ СЕМАНТИКАЛЫҚ ТАЛДАУЫ

Аннотация. Ақпараттық және смарт технологиялардың, жасанды интеллект жүйелерінің дамуына табиғи тілдерді өңдеу ғылыми зерттеу саласы үлкен ықпалын тигізіп жатыр. Мақалада семантикалық талдауға арналған қазіргі уақыттағы әдістер мен бағдарламалық тәсілдеріне жасалған жалпы шолу қамтылған. Жүргізілген зерттеулер негізінде мәтіндік ресурстарды семантикалық талдау үшін машиналық оқытуға негізделген әдіс көп қолданылатыны анықталды. Мақалада кілт сөздерді ескеру арқылы қазақ тіліндегі мәтінге семантикалық талдау жасаудың алгоритмі ұсынылған және модульдер жұмысы сипатталған. Белгілі бір ақпаратты белгілі бір атрибутқа сәйкестендіру алгоритмін жасағанда таңдау жасырын қабаты бар (100) нейрондық желіге тоқтады (NN). Бастапқы кезеңде мәтінді алдын ала өңделеді. Мәтінді алдын ала өңдеу үшін табиғи тілді өңдеуге арналған өзіміз жасаған модульдер қолданылды. Осы модульдерді қолданғаннан кейін дескриптор белгілері алынды. Содан кейін алынған мәліметтерді қолдана отырып, белгілер векторы құрылды. Құрылған белгілер векторы қазақ тілі үшін өзгертілген TF-IDF әдісімен анықталған белгілі бір кілт сөздермен салыстырылды. Екінші кезеңде нейрондық желі оқытылды. Сондай-ақ, берілген әдістің Python бағдарламалау тілінде орындалған бағдарламалық шешімі ұсынылған. Бағдарламалық жасақтаманы іске асыруда gensim, matplotlib, sklearn, numpy және т.б. сияқты кең тараған кітапханалар қолданылды. Модельді оқыту үшін келесі параметрлер орнатылды: мүмкіндік векторының өлшемі 100; сөйлемдегі ағымдағы және болжанатын сөз арасындағы аса үлкен арақашықтық 5; білімнің минималды деңгейі - 1; кесу жиілігі 4 сөзден тұрады. Сонымен қатар, модельді оқыту үшін қазақ тілінде 1 миллион сөйлемнен тұратын және SQL мәліметтер базасында орналасқан біртұтас қазақ корпусы дайындалды. Мәтінді модельмен өңдеген кезде KeyedVector-да model.wv модулінде сақталған сөз векторлары анықталады. Алынған сөз векторлары мәтін корпусындағы кілт сөздермен (сөз тіркестерімен) салыстырылады, бұл семантикалық атрибуттардың ықтимал мәнін әрі қарай пайдалану мақсатында жасалады. Нақты бір құжатқа арналған онтология нейрондық желінің жұмысы барысында пайда болатын семантикалық байланысты жұп жиынтығын қолдану арқылы жасалған. Жұмысымызда жүргізілген тәжірибе нәтижесі сөз жиынтығының графикалық түрінде көрсетілген. Ұсынылған тәсілдің жаңалығы – қазақ тіліндегі мәтін мағынасы жағынан жақын сөздерді анықтау. Бұл жұмыс машиналық аударма жүйесіндегі, ақпаратты іздеудегі, сонымен қатар талдау және өңдеу жүйесіндегі мәселелерді шешуге ықпал етеді.

Түйін сөздер: word2vec, модель, сөз, векторлық, көрініс, семантикалық, талдау, қазақ, тілі.

Д.Р. Рахимова[1,2], Ә.О. Тұрғанбаева[1]

[1]Казахский национальный университет имени аль-Фараби
[2]Институт информационных и вычислительных технологии, Алматы, Казахстан
E-mail: di.diva@mail.ru, turganbaeva.aliya@bk.ru

СЕМАНТИЧЕСКИЙ АНАЛИЗ КАЗАХСКОГО ЯЗЫКА НА ОСНОВЕ ПОДХОДА НЕЙРОННЫХ СЕТЕЙ

Аннотация. В данной работе представлен обзор существующих современных методов и программных подходов семантического анализа. На основе проделанных исследований выявлено, что для семантического анализа текстовых ресурсов наиболее применяется подход, основанный на машинном обучении. В данной статье представлен

разработанный алгоритм семантического анализа текста на казахском языке с учетом ключевых слов и описаны работы модулей. При разработке алгоритма для сопоставления определенной информации определенному атрибуту, выбор был остановлен на нейронной сети (НС) со скрытым слоем (100). Для начала выполняется предобработка текста. Для предобработки текста были использованы разработанные модули обработки естественного языка. После применения данных модулей были извлечены признаки нашего дескриптора. Затем с помощью извлеченных данных был построен вектор признаков. Построенный вектор признаков сопоставлялся с определенными ключевыми словами, определенный модифицированным методом TF-IDF для казахского языка. На втором этапе происходило обучение нейронной сети. В работе также представлено программное решение данного подхода, реализованного на языке программирования Python. В программной реализации были использованы известные библиотеки, такие как gensim, matplotlib, sklearn, numpy и т.д. Для обучения модели были заданы следующие параметры: Размерность векторов признаков составляет 100; Максимальное расстояние между текущим и предсказанным словом в предложении составляет 5; Минимальный уровень обучения 1; Пороговая частота среза 4 слов. Также для обучения модели был подготовлен одноязычный казахский корпус, который составляет 1 млн предложений на казахском языке и который находится в БД SQL. При обработке текста моделью выявляются вектора слов, которые хранятся в модуле model.wv в KeyedVectors. Полученные вектора слов также сопоставляются с ключевыми словами (словосочетаниями) из корпуса текстов с целью дальнейшего использования в качестве возможных значений семантических атрибутов сущностей. По набору семантически связанных пар слов строится онтология для конкретного документа, формирующаяся при работе нейронной сети. В работе представлены результаты проведенных экспериментов в графическом виде набора слов. Новизна предлагаемого подхода заключается во выявлении семантический близких слов по смыслу в текстах на казахском языке. Эта работа несет свой вклад в решение задач в системах машинного перевода, информационного поиска, а также в системах анализа и обработки на казахском языке.

Ключевые слова: модель, word2vec, векторное, представление, слов, семантический, анализ, казахский, язык.

Information about the authors:

Diana Rakhimova, Senior Lector at the Department of Information Systems, Al-Farabi Kazakh National University; Senior research of the Institute of Information and Computational Technologies, Almaty, Kazakhstan. di.diva@mail.ru. https://orcid.org/0000-0003-1427-198X;

Aliya Turganbayeva, Master of Technics and Technology, Teacher at the Department of Information Sytems, Al-Farabi Kazakh National University, Almaty, Kazakhstan. turganbaeva.aliya@bk.ru. https://orcid.org/0000-0001-9660-6928.

REFERENCES

[1] Pospelov D.A. Ten hotspots in artificial intelligence research // Intelligent Systems (MSU). 1996. T.1, No 1-4. P. 47–56.

[2] Alypansky G.A., Braslavsky P.I., Titov P.V. Formation of information queries to Internet search engines based on the thesaurus: a semantically oriented approach // Proceedings of the VIII Intern. conf. on electronic publications "EL-Pub2003". Novosibirsk, Akademgorodok, 2003. P. 269-270.

[3] Khairova, N., Petrasova, S., Mamyrbayev, O., Mukhsina, K. Open Information Extraction as Additional Source for Kazakh Ontology Generation // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020. P. 86-96.

[4] Mamyrbayev O., Toleu, A., Tolegen, G., Mekebayev, N. Neural architectures for gender detection and speaker identification // Cogent engineering, 2020. P. 1-7.

[5] Semantics // http://semantick.ru/: 14.07.2019.

[6] Tomita parser // http://api.yandex.ru/tomita/: 14.07.2019.

[7] In the foothills of semantics // http://dworq.com/: 29.05.2019.

[8] AI Data Analysis Technologies for Business // https://www.summarizebot.com/summarization_business.html: 27.05.2019.

[9] TextAnalyst ver. 2.0 – Program for personal analysis of texts // http://offext.ru/library/data/datakeeping/51.aspx: 19.04.2019.

[10] Galaktika-Zoom – analytical system for respectable clients // https://www.itweek.ru/themes/detail.php?ID=52215: 16.06.2019.

[11] Mamyrbayev, O. Zh; Shayakhmetova, A. S.; Seisenbekova, P. B. The methodology of creating an intellectual environment of increasing the competence of students based on a bayesian approach // News of the national academy of sciences of the Republic of Kazakhstan-series physico-mathematical, 2019. №4. (326). P. 50-58. **DOI:** 10.32014/2019.2518-1726.43

[12] Automatic text analysis technologies // http://nlp.isa.ru/: 26.04.2019.

[13] GitHub natasha // https://github.com/natasha: 26.04.2019.

[14] Manning Ch.D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, NY, USA, 2008. 210 p.

[15] Efficient Estimation of Word Representations in Vector Space // https://arxiv.org/pdf/1301.3781.pdf: 10.07.2018.

[16] Word2vec Parameter Learning Explained // https://arxiv.org/pdf/1411.2738.pdf: 10.07.2018.

[17] Texts in, Meaning out: neural language Models in semantic similarity tasks for russian // https://arxiv.org/ftp/arxiv/papers/1504/1504.08183.pdf: 20.04.2018.

[18] The Stanford Natural Language Processing Group // http://nlp.stanford.edu/software/CRF-NER.html: 19.08.2019.

Publication Ethics and Publication Malpractice
in the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethical guidelines for journal publication see **http://www.elsevier.com/publishingethics** and **http://www.elsevier.com/journal-authors/ethics**.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see **http://www.elsevier.com/postingpolicy**), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (**http://publicationethics.org/files/u2/New_Code.pdf**). To verify originality, your article may be checked by the Cross Check originality detection service **http://www.elsevier.com/editors/plagdetect**.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

(Правила оформления статьи для публикации в журнале смотреть на сайтах:

www:nauka-nanrk.kz

http://physics-mathematics.kz/index.php/en/archive

ISSN 2518-1726 (Online), ISSN 1991-346X (Print)

Редакторы: *М. С. Ахметова, Д. С. Аленов, А. Ахметова*
Верстка на компьютере *А.М. Кульгинбаевой*

Подписано в печать 22.09.2020.
Формат 60x881/8. Бумага офсетная. Печать – ризограф.
7,75 п.л. Тираж 300. Заказ 5.

Национальная академия наук РК
050010, Алматы, ул. Шевченко, 28, т. 272-13-18, 272-13-19