Using Word2vec Technique to Determine Semantic and Morphologic Similarity in Embedded Words of the Ukrainian Language

Larysa Savytska[a], Nataliya Vnukova[a], Iryna Bezugla[a], Vasyl Pyvovarov[b] and M. Turgut Sübay[c]

[a] *Simon Kuznets Kharkiv National University of Economics, Nauky av. 9a, Kharkiv, 61166, Ukraine*
[b] *Yaroslav Mudryi National Law University, Pushkinska str. 77, Kharkiv, 61024, Ukraine*
[c] *Piramit Danismanlik AS.: İstanbul, Kadıköy, Turkey*

Abstract

The study presents the word translation into vectors of real numbers (word embeddings), one of the most important topics in natural language processing. Word2vec is the latest techniques developed by Tomas Mikolov to study high quality vectors. The majority of studies on clustering the word vectors were made in English. Dmitry Chaplinsky has already counted and published vectors for the Ukrainian language by using LexVec, Word2vec and GloVe techniques, obtained from fiction, newswire and ubercorpus texts, for VESUM dictionary and other related NLP tools for the Ukrainian language. There was no research done on the vectors by using Word2vec technique to create Ukrainian corpus, obtained from Wikipedia dump as the main source. The collection contains more than two hundred and sixty one million words. The dictionary of words (unique words) obtained from the corpus is more than seven hundred and nine thousand. The research using machine technology Word2vec is of great practical importance to computerise many areas of linguistic analysis. The open-source Python programming language was used to obtain word vectors with Word2vec techniques and to calculate the cosine proximity of the vectors. In order to do machine learning with Word2vec techniques on Python, a resource containing open source licensed software libraries called "Gensim" was used. Calculations regarding the cosine affinities of the obtained vectors were made using "Gensim" libraries. The research examining the clustering of the word vectors obtained from the Ukrainian corpus was made considering the two sub-branches of linguistics, semantics and morphology (language morphology). Firstly, it was investigated how accurately the vectors are obtained from the Ukrainian corpus and how the words represent the cluster they belong to. Secondly, it was investigated how word vectors are clustered and associated respectively to the morphological features of the suffixes of the Ukrainian language.

Keywords 1

word2vec, NLP, cosine similarity, semantic relations, morphological (linguistics) relations, word vectors, word embedding, Ukrainian language

1. Introduction

Since the first years of computer science and technology, each technological step has offered humanity the possibility of storing larger amounts of data in smaller volumes and in cheaper way. Storage of large amounts of data, rapid analysis on data and data sharing has made computer science

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: larisa-savickaya@hotmail.com (L. Savytska); vnn@hneu.net (N. Vnukova); iryna.bezugla@hneu.net (I. Bezugla); v.pyvovarov@ukr.net (V. Pyvovarov); m.turguts@hotmail.com (M. T. Sübay);
ORCID: 0000-0002-9158-6304 (L. Savytska); 0000-0002-1354-4838 (N: Vnukova); 0000-0002-6285-2060 (I. Bezugla); 0000-0001-9642-3611 (V. Pyvovarov); 0000-0002-2967-694X (M. T. Sübay)

an important field of activity. Nowadays, computer technologies offer products that can be useful to all consumers of information services.

The large size of the data stored by computers brings the problem of finding the data quickly. The solution techniques of this problem constitute an important industrial field and are continuously being developed. Man-made analysis of data that needs content, meaning, emotion, commercial and similar needs is slow with growing data. The processing of big data with human control brings along a high cost problem. The solution of these problems in the line with industrial needs shows itself as software technologies that can perform automatic data analysis without human assistance. In today's world, automatic analysis is constantly being developed to meet the increasing industrial needs. Thanks to automatic analysis, information access, identifying people or objects from photographs, distinguishing the advertising contents of e-mails, analyzing sentiments in correspondence, translation between languages and many similar needs can be met. The research using machine technology Word2vec is of great practical importance to computerise many areas of linguistic analysis such as

- identifying semantic similarity of words and phrases
- automatic clustering of words according to the degree of their semantic closeness
- automatic generation of thesaurus and bilingual dictionaries
- resolving lexical ambiguity
- expanding queries due to associative connections
- defining the subject of the document
- clustering of documents for information retrieval
- extracting knowledge from texts
- constructing semantic maps of various subject areas
- modelling of periphrases
- determining the tone of the statement
- modelling of compatibility of word constraints.

2. Analysis of Publications

English mathematician Alan Turing asked a question "Can machines think like a human?" This proposal opened the idea of artificial intelligence and led to discussion [1] that artificial intelligence technologies can learn like humans and communicate with people.

Natural language processing started in 1950 with Alan Turing's publication "Computing machinery and intelligence" (known as "Turing test") [2, 3]. He is known as the father of theoretical computer science and artificial intelligence [1].

In 1986, David E. Rumelhart introduced the back propagation of the error to the world of artificial intelligence as a new learning technique in his study named "Learning representations by back-propagating errors" [4]. What is tried to be done in the new learning technique is the comparison of the vector that is known to be correct during the training phase and the vectors whose accuracy is estimated. The error value is obtained by making use of the difference between two vectors in comparison. Thanks to the obtained error value, corrections in the dimensional weights of the vectors are the basic principle on which the new technique is based. The process includes input vector, nonlinear hidden steps and output vector. The backward propagation of errors has made a significant contribution to the development of artificial neural networks.

In the error propagation technique, input data is given to the network as a vector and an output vector is produced. An error / loss amount (Error signal / Loss function) is calculated by comparing this output produced by the network with the training data whose result is known. Dimensional weights in the artificial neural network are updated according to the amount of error obtained. These updates are continued until the error amount drops to an appropriate level.

In the early 1990s, the error propagation technique began to be used in natural language processing. Yoshua Bengio, one of those who has made significant contributions to the development of natural language processing, has studied the Recurrent Neural Network (RNN) [5, 6, 7]. The RNN technique is based on the error propagation technique. The fact that statistics-based techniques are not suitable for practical use due to the long training period has prompted researchers to study more on

RNN [8]. With the adoption of the use of back propagation of error in natural language processing, discussions on learning techniques, deep learning and comparison of shallow learning techniques have become a broad area of research [9]. Tomas Mikolov stated in his studies on RNN [10] that statistics-based models do not reflect the meaning relationships of words well. He also stated that the vectors they obtained with the RNN model they developed can represent words more accurately.

Today, the technique of correcting the error backwards in natural language processing has become a current research area. The aim of this learning technique is to find the vector equivalents of the word in the multi-dimensional coordinate system consisting of real numbers. It is desired that the vectors corresponding to the words represent the word correctly (high quality vectors) in accordance with the structure of the language. The increase in vector quality increases their ability to reflect the versatile relationships found between words. Semantic results can be obtained by adding and subtracting quality vectors [11].

3. Using Word Embedding techniques in Research

The clusters and sub-clusters between the vectors obtained by machine learning are parallel in terms of the syntax of words, semantic and formal (structural) relationships. These relationships between words find wide application especially in industrial areas such as search engines. In natural language processing, the matching of words with vectors (finding word vectors) techniques are called word embeddings [12]. Vectors obtained with word embedding reflect the syntax and meaning relationships of the word without the need for human interaction [13]. One of the reasons for the development of word embedding techniques is that it shortens the machine learning training time. The shortening of the training period provides the opportunity to work with more vector dimensions and larger collections in practice. Being able to train machine learning with large corpus and more vector dimensions is shown among the important factors affecting the correct representation of words by vectors.

The three main factors that affect the correct representation of the word by vectors obtained by natural language processing machine learning are listed below.

1. The size of the corpus used in education is an important factor influencing the results obtained [8]. Increasing the corpus size causes more error correction operations on the weight of word vectors. It is possible to obtain more accurate vector values by increasing the corrections made on vector weights. The most important disadvantage of the growth of the corpus is that it extends the training period. This problem is possible to be overcome by developing new techniques [11, 14].
2. The size of the trained vectors is another important factor affecting the vector's ability to represent the word correctly. The vectors are represented in four dimensions. Each dimension of word vectors can be compared to the features related to supervised learning. It is observed that as the number of dimensions in unsupervised learning increases, similar to the features in supervised learning, the word vectors represent the word more accurately. In statistics-based techniques that are older than Word2vec techniques, the size of the vectors was given between 50 and 100 in order to keep the training time short. Y. Bengio stated in his publications that vector size growth naturally increases training times [7, 15]. The problem of long training periods has been reduced to a lesser extent in Word2vec. The number of vector sizes has been increased for Word2vec with the reduction of the time problem. It is recommended to specify the vector size between 300 and 1000 in Word2vec [11].
3. The number of neighbouring words used during training is one more important factor affecting the correct representation of the word by vectors. Increasing the number of neighbour words causes more error correction calculations to be performed on vectors. Increasing error corrections on vectors causes vectors to take more accurate weight values, but it has an increasing effect on training time. 5 to 10 neighbourhoods are recommended for Word2vec.

The size of the corpus, the size of the trained vectors and the number of adjacent words are important factors affecting vector quality. The study comparing the training times between the Word2vec techniques and the Neural Network Language Model (NNLM) technique, which Google

researcher Tomas Mikolov and his team have recently developed in their publications named "Efficient Estimation of Word Representations in Vector Space" [11], is given in the table below.

Table 1

Comparison of machine learning time between techniques (Although Word2vec technique has more vector dimensions than old techniques, it completes the training in a shorter time)

Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days x CPU cores]
			Semantic	Syntactic	Total	
NNLM	100	6B	34.2	64.5	50.8	14 x 180
CBOW	1000	6B	57.3	68.9	63.7	2 x 140
Skip-gram	1000	6B	66.1	65.1	65.6	2.5 x 125

The Neural Network Language Model shown in Table 1 is an older technique than Word2vec. The ensemble of consecutive words (CBOW) and prediction of neighbour words (Skip-gram) represent Word2vec algorithms. As can be seen in this table, in a collection composed of 6 billion words, the training process of NNLM technique is completed in 14 days with 180 central processing units (CPU cores). The first of two algorithms of the Word2vec completes the machine learning in 2 days with CBOW 140 CPU cores and the second with Skip-gram 125 CPU cores in 2.5 days.

4. Analysis of the Word2vec as a new technique in NLP

In the techniques used in natural language processing, the long duration of education has led researchers to reduce the time and thus to obtain more accurate word vectors by processing large collections. Google researcher Tomas Mikolov and his team, who conducted research on this issue, announced in 2013 that they have developed a new technique called Word2vec [11]. Word2vec technique is based on the error propagation technique similar to RNN. As it is seen in Table 1, the two algorithms in the newly developed Word2vec technique give better results than the NNLM technique. T. Mikolov states that the vectors obtained with Word2vec are clustered similar to the syntax and semantic relations of the words in the natural language.

Examples of syntax relationships between words in English are "great", "greater" or "easy", "easiest" word pairs. Similar to the syntax relations in these word pairs, T. Mikolov states that the vectors obtained with Word2vec are clustered.

T. Mikolov also shows with examples when the vectors obtained with Word2vec are clustered according to the semantic relations. As English word pairs, the word pair "Athens", "Greece" are in the semantic relation of country and capital. Similarly, the word pair "King" and "Queen" is in a semantic relation as expressions of nobility.

Semantic relationships between words cause the word vectors to cluster according to the semantic relationships of the words to which they belong. The correct representation of the word enables to obtain logical results. Semantic results can be obtained from the cosine similarities of the new vector obtained by adding and subtracting the vectors obtained with Word2vec. To give an example of semantic results that can be produced by arithmetic operations, the result vector obtained by replacing the gender feature in the word "King", which expresses nobility, is shown below [11].

('King') - ('Man') + ('Woman') result ('Queen')

In another example similar to this example, the semantic relationship between countries and their capitals is given below.

(England ') - (London') + ('Athens) result ('Greece)

Ukrainian is included in the East Slavic group of the Slavic branch of the Indo-European language family. English is also included in the Indo-European language family but belongs to the Germanic branch of languages. Ukrainian and English belong to the inflected languages. Their essential character is the division into languages of analytical and synthetical structure. The analytical structure presupposes the wide use of functional words, phonetic means and word-order for the expression of grammatical meanings. English is the language with analytical structure. The synthetical structure is

characterized by the greater role of the word forms which are created with the help of flexions and also word-forming suffixes and prefixes in the expression of grammatical meanings. Ukrainian belongs to the languages of synthetical structure. The examples above are given in English and it is possible to reproduce them, the corpus trained with Word2vec is mostly in English and a lot of work has been done on them.

Tomas Mikolov said "we should not expect only similar words to come close to each other, as there may be similarities in more than one way"[11]. In the example given in the same source, T. Mikolov stated that nouns can end with more than one suffix, and when searching for similar words, words that end with similar affixes can be reached. "... with the expectation that not only will similar words tend to be close to each other, but that words can have multiple degrees of similarity [16]. It has been observed earlier in the context of inflectional languages – for example, nouns can have multiple word endings, and if we search for similar words in a subspace of the original vector space, it is possible to find words that have similar endings [4, 17]".

In the literature review it was seen that some research was done on the vectors obtained by using Word2vec technique from the Turkish corpus, considering the meaning relation and formal features of the words [18]. Dmytro Chaplinskyi has already counted and published vectors for the Ukrainian language by using LexVec, Word2vec and GloVe techniques. It was a project to create corpus, obtained from fiction, newswire and ubercorpus texts, for VESUM dictionary and other related NLP tools for the Ukrainian language. [19]. There is no research done on the vectors obtained by using Word2vec technique to create Ukrainian corpus, obtained from Wikipedia dump as the main source. For this reason, vectors obtained from the Ukrainian corpus prepared for our research (with Word2vec) were subjected to consider the meaning relation and formal features of the words.

The words used in the natural language can have different meanings by establishing a contextual relationship with neighbouring words in their use in sentences. The contextual relationship between words leads to the establishment of multi-directional similarity relationships. These similarities may also occur according to the suffixes taken in inflected languages such as Ukrainian. This language is full of inflexion. Affixal morphemes in Ukrainian are mainly suffixes and sometimes prefixes. Even the number of suffixes considerably exceeds the number of prefixes. It was also stated by T. Mikolov that names can end with more than one suffix, and when searching for similar words in Word2vec, it is possible to reach words that end with similar suffixes [11].

5. Ukrainian Corpus trained with Word2vec: Tests and Results

The open-source Python programming language was used to obtain word vectors with Word2vec techniques and to calculate the cosine proximity of the vectors. In order to do machine learning with Word2vec techniques on Python, a resource containing open source licensed software libraries called "Gensim" was used [20, 21]. "Gensim" libraries were used to calculate the cosine similarities of the vectors. Continuous Bag-of-words (CBOW) algorithm is used for machine learning. The vector size is taken as three hundred and the window size as ten. The learning process is done in five cycles (EPOCH).

The cosine similarity values in the results are derived from the weight (dimension values) of the vector dimensions. Vector weights vary according to the size of the corpus used in education, the number of neighbouring words and the vector size. When a collection is trained twice with the same parameters, the cosine-like vector closest to the resulting vectors is expected to remain unchanged. Since the initial weights of the vectors are initialized by random value assignments, differences may occur in the result vectors and cosine similarity values. If the differences are large, it may be considered that the collection is not large enough. Increasing the number of cycles contributes to the vectors getting more accurate values.

Ukrainian corpus, obtained from Wikipedia dump as the main source, was used in the research. These are texts with general subject content. The collection contains more than two hundred and sixty one million words. The dictionary of words (unique words) obtained from the corpus is more than seven hundred and nine thousand. Such a big data gives an opportunity to conduct high quality semantic and morphologic analysis and Arithmetic operations of word vectors.

5.1. Semantic clustering of Ukrainian word vectors

Word vectors obtained with Word2vec over the general content Ukrainian corpus are clustered and related in terms of semantic relations of Ukrainian words.

The first example is the word "Яблуня". The first five word vectors with the closest cosine similarity to ('яблуня') vector are shown below.

[('груша', 0.8305871486663818),
('ожина', 0.8061103224754333),
('суниця', 0.8029876947402954),
('черешня', 0.7985619306564331),
('шовковиця', 0.797329306602478)]

The word "Яблуня" in the Academic Explanatory Dictionary of the Ukrainian language [22] is defined as

1. Noun, garden and forest fruit tree of the rose family with mostly spherical fruits.

It is clearly seen that among the vectors obtained by training from the Ukrainian corpus, there is the vector ('груша') as the closest cosine vector to the ('яблуня') vector. The word 'груша' in the Academic Explanatory Dictionary of the Ukrainian language is defined as

1. Noun, garden and forest fruit tree with dark green dense leaves and fruits, mostly in the shape of a rounded cone.
2. Noun, the fruit of this tree.
3. Noun, an object that has the shape of the fruit of this tree [22].

It can be clearly seen that the two words are in a semantic relationship.

Among the vectors obtained by training on the Ukrainian collection, there is the second closest cosine-like ('ожина') vector to ('яблуня') vector. The word 'ожина' in the Academic Explanatory Dictionary of the Ukrainian language is defined as

1. Noun, a perennial shrub or semi-shrub prickly plant with arcuate branches and edible fruits.
2. Noun, this plant's berries are black and grey [22].

It can be clearly seen that two words are in a paradigmatic relationship.

The other results obtained by training on the Ukrainian corpus are vectors belonging to lexical paradigm of the words representing the names of fruit trees, related to the meaning of the word "Яблуня".

As a result of training the word "Картопля", the first five word vectors with the closest cosine similarity to the vector ('картопля') are shown below.

[('квасоля', 0.8419397473335266),
('помідори', 0.8089007139205933),
('баклажани', 0.788599967956543),
('морква', 0.7878426313400269),
('кабачки', 0.787842631340434)]

The word "Картопля" in the Academic Explanatory Dictionary of the Ukrainian language is defined as

1. Noun, an annual herbaceous plant with edible tubers rich in starch.
2. Noun, tubers of this plant, used as food and as animal feed [22].

Among the vectors obtained by training on the Ukrainian corpus, the first closest cosine-like vector to ('картопля') vector is ('квасоля') vector. The word "Квасоля" in the Academic Explanatory Dictionary of the Ukrainian language is defined as

1. Noun, a garden herbaceous annual plant of the legume family, which has oval grains in the pods.
2. Noun, the fruits (oval grains) of this plant, which are eaten [22].

It is clearly seen that the two words are in a semantic affinity relationship. Referring to the other vectors such as 'помідори' and 'баклажани', they belong to lexical paradigm of the words representing the names of vegetables, related to the meaning of the word 'картопля'.

As a result of training the word "Харків", the city name, the first five word vectors with the closest cosine similarity to the vector ('харків') are shown below.

[('київ', 0.6517883539199829),
('дніпро', 0.5908591747283936),
('полтава', 0.5591470003128052),
('донецьк', 0.5527917742729187),
('львів', 0.5408258438110352)]

The word "Харків" in the Universal dictionary-encyclopaedia is defined as

Regional centre of Ukraine, in the place of convergence for the Kharkiv, Lopan and Udy rivers (Severskyi Donets Basin); 1.6 million people (second after Kyiv in terms of population in Ukraine) [23].

Among the vectors obtained by training on the Ukrainian corpus, the first closest cosine-like vector to ('харків') vector is ('київ') vector. The word "Київ" in the Universal dictionary-encyclopaedia is defined as

The capital of Ukraine, the city of state subordination, the centre of the region and Kyiv-Sviatoshynskyi district; on both sides of the Dnieper River, in its middle course, below the confluence of the left tributary of the Desna, the rivers Lybid, Syrets, Vita (right tributaries of the Dnieper), Gorenka, and Nivka (a tributary of the Irpen) also flow along the Kyiv; 2.6 million people [23].

It is clearly seen that the two words are in a semantic affinity relationship. Referring to the other vectors such as 'дніпро' and 'полтава', they belong to lexical paradigm of the words representing the other cities names in Ukraine, related to the meaning of the word 'Харків'.

As a result of training the word "Микола", a proper name, the first five word vectors with the closest cosine similarity to the vector ('Микола') are shown below.

[('михайло', 0.8178503513336182),
('олександр', 0.7979997396469116),
('василь', 0.7977378368377686),
('федір', 0.7911434173583984),
('петро', 0.7876654267311096)]

The word "Микола" is used as a neutral man’s name in Ukrainian. Among the vectors obtained by training from the Ukrainian corpus the closest cosine-like ('микола') vector is the ('михайло') vector. The word "Михайло" is also used as a neutral male name in Ukrainian language. When the vectors that similar to the closest cosine to the vector ('михайло') are examined, the vectors belonging to words / proper names representing male names in similar usage with the use of the word "Михайло" are investigated. From the results, it is clearly seen that there is a semantic cluster related to the usage area of the word "Микола".

As a result of training the word "Леся", a proper name, the first five word vectors with the closest cosine similarity to the vector ("Леся") are shown below.

[('оксана', 0.6321463584899902),
('соломія', 0.5867637395858765),
('орися', 0.5651060342788696),
('михайлина', 0.5579890012741089),
('наталя', 0.5555435419082642)]

The word "Леся" is used as a specific woman’s name in Ukrainian. Among the vectors obtained by training from the Ukrainian corpus the closest cosine-like ('леся') vector is the ('оксана') vector. The word "Оксана" is also used as a woman’s name in Ukrainian language. When the vectors that similar to the closest cosine to the vector ('оксана') are examined, the vectors belonging to words /

proper names representing female names in similar usage with the use of the word "Оксана" are investigated. From the results, it is clearly seen that there is a semantic cluster related to the usage area of the word "Леся".

According to the results, the words "Микола" and "Леся" differ in their gender characteristics but they are in a semantic cluster related to proper names either neutral or specific.

As a result of training the word "Білий", qualitative adjective, the first five word vectors with the closest cosine similarity to the vector ('білий') are shown below.

[('чорний', 0.758683443069458),
('блакитний', 0.6943730711936951),
('жовтий', 0.6647096872329712),
('синій', 0.6603621244430542),
('червоний', 0.6490797996520996)]

The word "Білий" in the Academic Explanatory Dictionary of the Ukrainian language is defined as

1. Adjective. It has the colour of chalk, milk, snow; the opposite is black [22].

Among the vectors obtained by training from the Ukrainian corpus, the first closest cosine-like vector to ('білий') vector is ('чорний') vector. The word "Чорний" in the Academic Explanatory Dictionary of the Ukrainian languageis defined as

1. Adjective. Colour of soot, coal, and the darkest; opposite white [22].

It is clearly seen that two words are in a semantic relationship, even they are antonyms. Referring to the other vectors such as 'блакитний' and 'жовтий', they belong to lexical paradigm of the words representing the names of colours, related to the meaning of the word 'білий'.

According to the results obtained by training the Ukrainian corpus it is proved that the vectors are clustered and related in terms of semantic relations of Ukrainian words.

5.2. Arithmetic operations of word vectors and semantic relationships between words

New vectors can be obtained as a result of adding and subtracting (arithmetic operations) the word vectors obtained from the Ukrainian corpus.

The first example is similar to the English example, showed by T. Mikolov [11] obtained from the English corpus when the cosine analogues of the new vector are obtained by adding and subtracting the vectors.

(‘king’) - (‘man’) + (‘woman’) = (‘queen’)

The first five word vectors with the closest cosine similarity to the result vector of ('король') - ('чоловік') + ('жінка') operation are shown below.

Cosine similarity results of ('король') - ('чоловік') + ('жінка') operation result vector:

[('королева', 0.6145955324172974),
('принцеса', 0.4626465737819671 6),
('правителька', 0.4591668248176574 7),
('корона', 0.4428777694702148 4),
('королевою', 0.4254585504531860 4)]

The result obtained from the Ukrainian corpus is similar to the result obtained from the English corpus. The word "Королева" is the Ukrainian equivalent to the word "Queen". The vector was found to be the closest cosine-like vector to the result from the process.

The ('король') - ('чоловік') + ('жінка') operation is the replacement of the gender feature in the word "Король", which expresses nobility. In terms of the word meaning, the result of the process is the word "Королева". It is seen that the word meaning is compatible with the result of adding and subtracting the vectors. The word "Королева" is defined as "the wife of the king or the woman who rules the kingdom" in the Academic Explanatory Dictionary of the Ukrainian language [22].

The first five word vectors with the closest cosine similarity to the result vector of ('англія') - ('лондон') + ('київ') operation are shown below.

Cosine similarity results of ('англія') - ('лондон') + ('київ') operation result vector:

[('україна', 0.5563449859619141),
('одеса', 0.4825827181339264),
('чернігів', 0.4495120048522949),
('дніпропетровськ', 0.4358903467655182),
('харків', 0.40171462297439575)]

The ('англія') - ('лондон') + ('київ') operation is the transaction of the relationship between countries and cities (or their capitals). According to the results of operations with the word "Україна", it is seen that the word meanings are compatible with the result of adding and subtracting vectors.

The first five word vectors with the closest cosine similarity to the result vector of ('банк') - ('золото') + ('кредит') operation are shown below.

Cosine similarity results of ('банк') - ('золото') + ('кредит') operation result vector:

[('капітал', 0.5785767436027527),
('кредитний', 0.5400320291519165),
('інвестор', 0.5385503172874451),
('банкінг', 0.5157139301300049),
('позичальник', 0.5080622434616089)]

The word "Банк" in the Great explanatory dictionary of the modern Ukrainian language is defined as

1. A credit and financial institution that concentrates funds and investments, provides loans, makes cash settlements between enterprises or individuals, and regulates money circulation in the country, including the issuance of new money [24].

The word "Золото" in the Great explanatory dictionary of the modern Ukrainian language is defined as

1. The simple substance of the chemical element Aurum is yellow, soft, malleable metal.
2. Golden things; expensive gold-woven clothes, etc.
3. Gold coins, money, etc.
4. About something very valuable, beautiful or about someone worthy of respect.
5. About gold medal (gold medals) for a victory, the first place in sports competitions, on competition, etc. [24].

The word "Кредит" in the Great explanatory dictionary of the modern Ukrainian language is defined as

1. Lending of material values, money; loan.
2. Budget amounts in the outlay, within which the costs of something are allowed [24].

The semantic result obtained from the ('банк') - ('золото') + ('кредит') operation is the result of redefining investments and loans into capital by making profits and surpluses. According to the results of operations with the word "Капітал", it is seen that the word meanings are compatible with the result of adding and subtracting vectors. Referring to the other vectors such as 'кредитний', 'інвестор', 'банкінг', 'позичальник' they are in the semantic affinity relationship.

The first five word vectors with the closest cosine similarity to the result vector of ('спорт') - ('гімнастика') + ('баскетбол') operation are shown below.

Cosine similarity results of ('спорт') - ('гімнастика') + ('баскетбол') operation result vector:

[('бейсбол', 0.6380308866500854),
('гольф', 0.5946487784385681),
('футбол', 0.5910741090774536),
('крикет', 0.5801275372505188),
('регбі', 0.5448272824287415)]

The word "Спорт" in the Academic Explanatory Dictionary of the Ukrainian language is defined as

1. Physical exercises (gymnastics, athletics, sports, tourism, etc.), which aim to develop and strengthen the body or mind [22].

The ('спорт') - ('гімнастика') + ('баскетбол') operation is the transaction of exchange of two sports branches in the vector operation. When the results are evaluated, the result vectors are compatible in terms of meaning relations of the words they belong to.

The clusters between the vectors obtained from the Ukrainian corpus, exampled above, were examined considering the semantic relationship between the words they belong to. It is clearly seen that semantic relations between Ukrainian words build clusters in the vectors. It is proved that semantic results obtained by addition and subtraction on vectors obtained from the English corpus can be also obtained from the Ukrainian corpus.

5.3. Formal clustering of Ukrainian word vectors

Ukrainian is a language of inflexion. Affixal morphemes in Ukrainian are mainly suffixes and sometimes prefixes. Even the number of suffixes considerably exceeds the number of prefixes. The general feature of Ukrainian language is that the word roots are kept constant and the constructions and inflections, which have various functions, are added to the roots. By adding different suffixes to the roots of the word, new words are derived and the vocabulary of the language is formed in this way.

When searching similar words by using word vectors, it is seen that words ending with similar suffixes can be reached [11]. The word vectors obtained from Ukrainian corpus (with Word2vec) are clustered and related according to the Ukrainian-specific suffixes.

The first word to be examined is the word "Ходити". The first five word vectors with the closest cosine similarity to vector ('ходити') are shown below.

The cosine similarity results of vector ('ходити'):

[('бігати', 0.7297786474227905),
('їздити', 0.7247925996780396),
('лазити', 0.7013007402420044),
('сидіти', 0.7005398273468018),
('гуляти', 0.697922945022583)]

The first five words clustered like a cosine are vectors belonging to the verbs in the form of infinitive. The clustering of word vectors is related to the formal feature infinitive suffix "-ти".

The word "Ходив" is derived by taking the past tense singular affix "-в" to the root "ход". The first five word vectors with the closest cosine similarity to vector ('ходив') are shown below.

The cosine similarity results of vector ('ходив'):

[('бігав', 0.6834566593170166),
('водив', 0.678192138671875),
('вирушав', 0.6492938995361328),
('приїжджав', 0.6308779716491699),
('гуляв', 0.6269496083259583)]

The first five words "бігав", "водив", "вирушав", "приїжджав", "гуляв" are vectors belonging to the verbs in the form of the past tense singular derived by adding the past tense singular affix to the verb roots. The clustering of word vectors is related to the formal feature the past tense singular affix "-в".

The word "Прачка" is derived by taking the feminine suffix "-к" to the root "прач". The job name is derived from the verb describing the process. The top five word vectors with the closest cosine similarity to vector ('прачка') are shown below.

The cosine similarity results of the vector ('прачка'):

[('кухарка', 0.6516879796981812),
('економка', 0.6106237173080444),
('хазяйка', 0.6105634570121765),
('покоївка', 0.5945062637329102),
('нянька', 0.581095814704895)]

The first five words clustered like a cosine are vectors belonging to the nouns describing the job name. The clustering of word vectors is related to the formal feature the feminine noun suffix "-к".

The word "Яблуня" was discussed while analyzing the semantic relationships between vectors. For the word "Яблуневий" in the sentence "Мені подобається яблуневий пиріг" the first five word vectors with the closest cosine similarity to the vector ("Яблуневий") are shown below.

The cosine similarity results of the vector ('яблуневий'):

[('вишневий', 0.6773509979248047),
('квітучий', 0.6477996110916138),
('райський', 0.6233205199241638),
('фруктовий', 0.6139903664588928),
('виноградний', 0.6060866117477417)]

The word "Яблуневий" in the sentence "Мені подобається яблуневий пиріг" is derived by adding the suffix "-ев" which makes the word an adjective with the word meaning apple tree. Among the vectors obtained by training from the Ukrainian corpus the closest cosine-like vector is 'вишневий'. The clustering of word vectors is related to the formal feature adjective suffix "-ев".

The clusters between the vectors obtained from the Ukrainian corpus were examined considering the formal relationship between the words. It is proved that the word vectors obtained from Ukrainian corpus are clustered and related according to the Ukrainian-specific suffixes.

5.4. Arithmetic operations of word vectors and formal relation between words

New vectors can be also obtained as a result of adding and subtracting (arithmetic operations) the word vectors obtained from the Ukrainian corpus by examining the formal clustering.

The first five word vectors with the closest cosine similarity to the result vector of ('квіти') - ('квітка') + ('яблуко') operation are shown below.

Cosine similarity results of ('квіти') - ('квітка') + ('яблуко') operation result vector:

[('яблука', 0.5736010074615479),
('руно', 0.4630397856235504),
('перо', 0.4570971727371216),
('теля', 0.44505059719085693),
('курча', 0.4449284076690674)]

According to the results of operations ('квіти') - ('квітка') + ('яблуко'), the first vector 'яблука', clustered like a cosine, is a vector belongs to the formal feature plural form of nouns. The clustering of word vectors 'руно', 'перо', 'теля', 'курча' is related to the formal feature neuter gender.

The first five word vectors with the closest cosine similarity to the result vector of ('олівці') - ('олівець') + ('ручка') operation are shown below.

Cosine similarity results of ('олівці') - ('олівець') + ('ручка') operation result vector:

[('ручки', 0.6751487255096436),
('фольга', 0.6358448266983032),
('пластмаси', 0.6101338863372803),
('накладки', 0.6093952655792236),
('упаковка', 0.5968321561813354)]

According to the results of operations ('олівці') - ('олівець') + ('ручка'), the vectors 'ручки', 'пластмаси', 'накладки', clustered like a cosine, is a vector belongs to the formal feature plural form of nouns. The clustering of word vectors 'фольга', 'упаковка' is related to the formal feature noun, feminine gender.

The first five word vectors with the closest cosine similarity to the result vector of ('яблуневий') - ('яблуня') + ('вишня') operation are shown below.

Cosine similarity results of ('яблуневий') - ('яблуня') + ('вишня') operation result vector:

[('вишневий', 0.5016539096832275),
('гетсиманський', 0.48334306478500366),
('веселий', 0.45027756690979004),
('грушка', 0.4389912188053131),
('гефсиманський', 0.43634486198425293)]

According to the results of operations ('яблуневий') - ('яблуня') + ('вишня'), the vectors 'вишневий', ‘гетсиманський’, 'веселий', 'гефсиманський', clustered like a cosine, is a vector belongs to the formal feature adjective, masculine gender.

The clusters between the vectors obtained from the Ukrainian corpus were examined considering the formal relationship between the words. It is proved that the formal results obtained by addition and subtraction on vectors are clustered and related according to the Ukrainian-specific suffixes.

6. Conclusions

The research using machine technology Word2vec is of great practical importance to computerise many areas of linguistic analysis. The results obtained regarding to the clustering of word vectors proved that there is an affinity in the construction of words in both semantic and morphologic similarity, that indicates a high structural level of construction of the Ukrainian language.

Clustering of vectors obtained from the Ukrainian corpus can be in a semantic affinity relationship with the words they belong to, form relations or both. Clusters can go down to sub-breakdowns.

In the cosine similarities of the vector belonging to the word "Яблуня", the word "Яблуня" is clustered with vectors belonging to lexical paradigm of the words representing the names of fruit trees.

In the cosine similarities of the vector belonging to the word "Картопля", the word "Картопля" is clustered with vectors belonging to lexical paradigm of the words representing the names of vegetables.

In the cosine similarities of the vector belonging to the word "Харків", the words "Дніпро" and "Полтава" represent the other cities names in Ukraine and belong to lexical paradigm of the words related to the meaning of the word 'Харків'.

In the cosine similarities of the vector belonging to the word "Микола", the neutral male name used in Ukrainian is clustered together with other neutral male names.

In the cosine similarities of the vector belonging to the word "Леся", the specific female name used in Ukrainian is clustered together with other specific female names.

In clustering proper names, separate clustering of male and female names is an example of sub-break.

In the cosine similarities of the vector belonging to the word "Білий", the first vector 'чорний' is in a semantic antonyms relationship. Referring to the other vectors such as 'блакитний' and 'жовтий', they belong to lexical paradigm of the words representing the names of colours, related to the meaning of the word 'білий'.

New vectors can be obtained as a result of adding and subtracting (arithmetic operations) the word vectors obtained from the Ukrainian corpus. The cosine similarities of the vectors obtained by the process were examined in terms of their compatibility with the meaning of the process. It is proved that the semantic results that can be obtained by addition and subtraction on vectors obtained from the English corpus can be also obtained from the Ukrainian corpus.

The vector obtained as a result of ('король') - ('чоловік') + ('жінка') operation is ('королева') the first vector among the cosine-like vectors. The process and result vectors are compatible with the semantic relationships of the words they belong to.

The vector obtained as a result of ('англія') - ('лондон') + ('київ') operation is ('україна') the first vector among the cosine-like vectors. It is the transaction of the relationship between countries and cities (or their capitals). The process and result vectors are compatible with the semantic relationships of the words they belong to.

The vector obtained as a result of ('банк') - ('золото') + ('кредит') operation is 'капітал' the first vector among the cosine-like vectors. It is the result of redefining investments and loans into capital by making profits and surpluses. According to the results of operations with the word "Капітал", the word meanings are compatible with the result of adding and subtracting vectors. Referring to the other vectors such as 'кредитний', 'інвестор', 'банкінг', 'позичальник' they are in the semantic affinity relationship.

The vector obtained as a result of ('спорт') - ('гімнастика') + ('баскетбол') operation is 'бейсбол'. It is the transaction of exchange of two sports branches in the vector operation. When the results are evaluated, the result vectors are compatible in terms of meaning relations of the words they belong to.

Considering the morphological properties of the words, the vectors can be clustered according to the suffixes they take. Clustering according to suffixes can also include semantic relationships of words.

In the cosine similarities of the vector belonging to the word "Ходити", the clustering of word vectors is related to the formal feature infinitive suffix "-ти".

In the cosine similarities of the vector belonging to the word "Ходив", the clustering of word vectors is related to the formal feature the past tense singular affix "-в".

In the cosine similarities of the vector belonging to the word "Прачка", clustering of word vectors is related to the formal feature the feminine noun suffix describing the job name "-к".

In the cosine similarities of the vector belonging to the word "Яблуневий", the clustering of word vectors is related to the formal feature adjective suffix "-ев".

New vectors can be also obtained as a result of adding and subtracting (arithmetic operations) the word vectors obtained from the Ukrainian corpus by examining the formal clustering.

According to the results of operations ('квіти') - ('квітка') + ('яблуко'), the first vector 'яблука', clustered like a cosine, is a vector belongs to the formal feature plural form of nouns. The clustering of word vectors 'руно', 'перо', 'теля', 'курча' is related to the formal feature neuter gender.

According to the results of operations ('олівці') - ('олівець') + ('ручка'), the vectors 'ручки', 'пластмаси', 'накладки', clustered like a cosine, is a vector belongs to the formal feature plural form of nouns. The clustering of word vectors 'фольга', 'упаковка' is related to the formal feature noun, feminine gender.

According to the results of operations ('яблуневий') - ('яблуня') + ('вишня'), the vectors 'вишневий', ‘гетсиманський’, 'веселий', 'гефсиманський', clustered like a cosine, is a vector belongs to the formal feature adjective, masculine gender.

More detailed analysis to computerise such areas of linguistic as constructing semantic maps of various subject areas and expanding queries due to associative connections will be the subjects of our further research.

7. References

[1] Britannica, The Editors of Encyclopaedia, Turing test, Artificial intelligence, 2020. URL: https://www.britannica.com/technology/Turing-test

[2] A. M. Turing, Computing Machinery and Intelligence, Mind (1950) 433–460. doi:10.1093/mind/lix.236.433

[3] T. Mendès France, Turing et son test : une alchimie contemporaine ? Notes sur les critiques des scientifiques des années 90, Quaderni (1996) 41–46. doi:10.3406/quad.1996.1953

[4] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, Nature (1986) 533–536. doi:10.1038/323533a0

[5] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Transactions on Neural Networks (1994) 157–166. doi: 10.1109/72.279181
[6] S. El Hihi, Y. Bengio, Hierarchical recurrent neural networks for long-term dependencies, in: Proceedings of the 8th International Conference on Neural Information Processing Systems, NIPS'95, MIT Press, Cambridge, MA, USA, 1995, pp. 493–499. doi:10.5555/2998828.2998898
[7] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, Journal of Machine Learning Research (2003) 1137–1155. doi:10.1162/153244303322533223
[8] T. Mikolov, M. Karafiat, L. Burget, J Cernocky, S. Khudanpur, Recurrent neural network based language model, in: Proceedings of Interspeech-2010, International Speech Communication Association, Makuhari, Chiba, JP, 2010, pp. 1045–1048.
[9] Y. Bengio, Y. Lecun, Scaling learning algorithms towards AI, in: L. Bottou, O. Chapelle, D. DeCoste, J. Weston (Eds), Large-scale kernel machines, Mit Press, Cambridge, Mass, 2007. doi:10.7551/mitpress/7496.003.0016
[10] T. Mikolov, S. Kombrink, L. Burget, J. Černocký, S. Khudanpur, Extensions of Recurrent Neural Network Language Model, in: Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, IEEE Signal Processing Society, Praha, CZ, 2011, pp. 5528–5531. doi:10.1109/ICASSP.2011.5947611
[11] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, in: Proceedings of Workshop at ICLR 2013, Computation and Language Scottsdale, Arizona, USA, 2013. arXiv:1301.3781v3
[12] R. Lebret, R. Collobert, Word Embeddings through Hellinger PCA, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 482–490. doi:10.3115/v1/E14-1051
[13] Y. Chen, B. Perozzi, R. Al-Rfou, S. Skiena, The Expressive Power of Word Embeddings, in: Proceedings of the 30 th International Conference on Machine Learning, ICML 2013, Atlanta, Georgia, USA, 2013. arXiv:1301.3226
[14] T. Brants, A. Popat, P. Xu, F. Och, J. Dean, Large Language Models in Machine Translation, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language, EMNLP-CoNLL, Association for Computational Linguistics, 2007, pp. 858–867.
[15] Y. Bengio, J.-S. Senecal, Quick Training of Probabilistic Neural Nets by Importance Sampling, in: Proceedings of AISTATS 2003. Society for Artificial Intelligence and Statistics, Florida, USA, 2003.
[16] J. Weizenbaum, ELIZA - a computer program for the study of natural language communication between man and machine, Communications of the ACM, (1966) 36–45. doi:10.1145/365153.365168
[17] T. Mikolov, Q.V. Le, I. Sutskever, Exploiting Similarities among Languages for Machine Translation; Computing Research Repository (CoRR), 2013. arXiv:1309.4168
[18] M. T. Sübay, Türkçe kelime vektörlerinde görülen anlamsal ve biçimsel yakınlaşmalar [The semantic and morphologic similarity in Turkish word embeddings]. Maltepe Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul, 2019. URL: https://hdl.handle.net/20.500.12415/2733
[19] A. Rysin, V Starko, D. Chaplynskyi, Slovnyk VESUM ta inshi poviazani zasoby NLP dlia ukrainskoi movy [VESUM dictionary and other related NLP tools for the Ukrainian language], 2007. URL: https://r2u.org.ua/articles/vesum
[20] Gensim: topic modelling for humans. Corpus from a Wikipedia dump. URL: https://radimrehurek.com/gensim/corpora/wikicorpus.html
[21] Gensim: topic modelling for humans. Word2vec embeddings. URL: https://radimrehurek.com/gensim/models/word2vec.html
[22] Slovnyk ukrainskoi movy. Akademichnyi tlumachnyi slovnyk [Academic Explanatory Dictionary of the Ukrainian language], 2018. URL: http://sum.in.ua/
[23] USE (Universalnyi slovnyk-entsyklopediia) [UDE (Universal Dictionary-Encyclopedia)], Slovopedia, 2007. URL: http://slovopedia.org.ua/29/53392-0.html
[24] Velykyi tlumachnyi slovnyk (VTS) suchasnoi ukrainskoi movy [Great explanatory dictionary of the modern Ukrainian language], Slovopedia, 2007. URL: http://slovopedia.org.ua/93/53393/828300.html, http://slovopedia.org.ua/93/53399/882529.html