*МРНТИ 20.23.19*

*D.A.AYAZBAYEV [1], A.MURATKYZY[1], R.Z.ZHUMALIYEVA[1]*

*[1]Suleyman Demirel University (Kaskelen, Kazakhstan),*
*Dauren.Ayazbayev@sdu.edu.kz; akbotamuratkyzy19@gmail.com;*
*rzhresearch2019@gmail.com; https://doi.org/10.51889/2020-3.2077-6861.03*

# WORD EMBEDDING IN TEACHING RESEARCH WRITING

*Abstract*

In research writing as a scientific basis in education, some mistakes related to choosing appropriate term or definition take place. This research discusses the solving such problem by using word embedding in the Research Writing discipline. This is a word representation form, where one word has a vector and its coordinates. The words with close meaning have similar direction, showing lexical compatibility. To calculate lexical relations, the cosine of the angle between two words' vectors are considered. Value of highly compatible word combinations is equal to 1. On the other hand, lexically incompatible words should approximately have value -1.

To test the system the text of the Constitution of the Republic of Kazakhstan was used. Particularly, words which are not related to meaning of article of the Constitution were inserted, and the system had to identify that inserted words. The system for some words showed high accuracy, however some words showed low accuracy. It is suggested that such factor was because even inserted words were not related in meaning, they could be lexically compatible with their neigh bors.

This research is carried out within the framework of the Ministry of Education and Science of Republic of Kazakhstan grant project "Developing and implementing the innovative competency-based model of multilingual IT specialist in the course of national education system modernization".

*Keywords:* lexical compatibility of words; neural network; Skip-gram model; vector of word.

**Introduction.** The process of writing research papers plays a great role in education, forming scientific basis for a speciality. However, sometimes in research writing, there are mistakes in choosing appropriate term or definition. Such problem can be solved by checking the terms for being suitable in a definite context. One of the methods is word embedding, which is a form of word representation. While processing, word embedding convert words into vectors. Each one of the vectors has own coordinates, which helps to identify it. So, words with close meanings should be in about the same direction. In addition, when word embedding determines the coordinates of a word vector, the lexical combination with another word is considered. This helps to choose the most appropriate term regarding meaning and context. It is necessary to highlight that it can be taught in Research Writing discipline and improve the scientific works of the students

**Literature review.** Word embedding can be used in many different areas. For example, Bondareva and Lagerev used word embedding to get people's opinions [1, PP.10-15]. Furthermore, Wohlgenannt and others used such process to determine the interaction of the characters, i.e. the social network, in the books [10, PP.18-25]. Mikolov suggested to use word embedding in machine translation, saying that learned structure of word relationships in one language often correlates with another language [7]. Further research of Kusner and others implies implementing such method in identifying not just similarity between words, but also documents [4].

Analysing the last research and experiments, it can be made a block diagram of the word embedding process, which is shown in Figure 1.

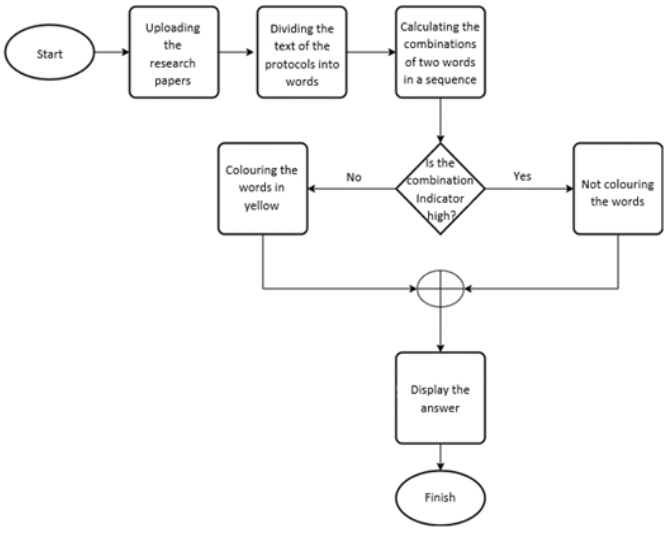


Figure 1. **Block diagram of the project**

As shown in Figure 1, the system begins with the uploading the research papers, so system has a dictionary of word vectors. If the vectors of two words are known, it gives opportunity to calculate the lexical combination of those words.

**Materials and methods.** There are several models that turn a word into a vector. For example: Skip-gram, continuous bag of words, GloVe. In this research, the Skip-gram model was used for determining and converting the vector of a word by the help of a neural network [5]. The Skip-gram model determines the probability of occurrence of a given word within the context of neighbouring words. To determine the vector of a word in such network, it is needed to go through following steps: [2;3;6]

1) To determine the word of the vector, extract the sentences in which it occurs from the corpus. Then repetitive words from those sentences should be removed. The rest of the words forms the input layer of the neural network. The structure of the neural network is shown in Figure 2.

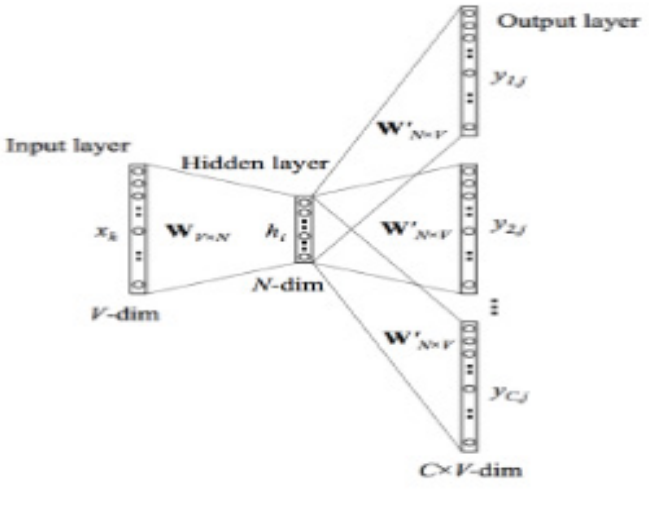


Figure 2. **The structure of Skip-gram's neural network**

Here:

x – are the neurons of the input layer.

W – weights between neurons.

h – neurons of the latent layer.

y – neurons of the output layer.

V – number of different words.

C is the number of neighbours of the word, the vector of which is needed to determine (window size).

One neuron corresponds to each word in the input layer.

2) All neurons except the neuron of the word selected for the search receive a value of 0. The neuron of the word selected for the search is equal to 1.

3) In a neural network, all weights are assigned between 0 and 1. The weights can be different in the input and latent layers, latent and output layers.

4) Enter the values of all neurons in the input layer and multiply it by the weights in the latent layer.

$$h=x^T W \quad (1)$$

5) The values of the neurons in the output layer are determined by the following formula:

$$u=h W^T \quad (2)$$

W – the weight between the latent layer and the output layer.

6) The values of the neurons of the output layer are converted to the probabilities with the softmax function. The following formula is used for this:

Here:

*wc,j* – the j-word in the c-context in the output layer.

*wo,c* – the c-word in the output layer.

*wi* – the word of the vector to be defined in the input layer.

yc,j – the occurrence probability of the j-word's neighbour in the c-context.

In a neural network, the number of neighbours of the word, the vector of which is needed to determine, defines the number of the context. Each context corresponds to one neighbour.

7) For each neuron in the output layer, the prediction error is calculated:

If the j-word in the c-context is a neighbour of the c-context, tc,j is equal to 1. It is 0 in rest cases.

8) All errors in the words of the output layer are added:

$$EI_j = \sum_{c=1}^{C} e_{c,j} \quad (5)$$

Here:

C – number of the context.

9) All weights in the neural network are updated with the following formula:

□w_(i,j)□^((new))=□w_(i,j)□^((old))-α □EI□_(j ) h_i (6)

Here:

α – learning rate.

*wi,j*(new) – new weight.

*wi,j*(old) – old weight.

*hi* – the value of the neuron of the latent layer.

10) Repeat steps 4-9 until the neural network error is low.

In this research, the cosine of the angle between the vectors of the two words was calculated to determine the lexical combination of the two words. The closer the meanings of the two words are [8;9], or the larger the lexical combinations, the larger the value of the cosine. The cosine of the angle between the two vectors is calculated by the following formula:

cos□□(α)=□ (a_1 b_1+a_2 b_2+...+a_n b_n)/(√(a_1^(2 ) □+a□_2^2 □…+a□_n^2 ) √(b_1^(2 ) □+b□_2^2 □…+b□_n^2 )) (7)

Here:
*n* – the size of the vector.
In this research, n was equal to 100.

**Discussion and results.** For checking the lexical combinations of words, the Constitution of the Republic of Kazakhstan was chosen as a text with stable word combinations. When writing a new word between the words of subparagraph 1 of Article 90, the system had to find that new word. To find a new word, the cosines between the vectors of the words were calculated. There were words to the right and left of the new word. If the cosines of the words entered to the left or right of the new word were smaller than the cosines of the other words, the system was considered to have found the new word entered. Table shows the accuracy of the system.

Table 1

**System accuracy**

| *Word* | Accuracy |
|---|---|
| *Green* | 57.14% |
| *Cow* | 100% |
| *Wolf* | 100% |
| *Pen* | 71.43% |
| *Computer* | 57.14% |
| *Astronaut* | 28.57% |
| *Automobile* | 14.29% |
| *Airplane* | 71.43% |
| *Iron* | 71.43% |
| *Aluminum* | 85.71% |

To calculate the accuracy of the system, a new word was placed between the different words in subparagraph 1 of Article 90. As shown in Table 1, the system returned with different accuracy. Only the words cow and wolf showed 100% accuracy. Because these words are not lexical combinations with the words of subparagraph 1 of Article 90 (for example: *Republican cow, official cow*). However, for some words, the accuracy of the system was low. This is because those words can be lexically combined with the word on the right or left, even if their meaning does not correspond to subparagraph 1 of Article 90. For example, the word airplane can be lexically combined with the word *former,* the word astronaut with the *official.*

**Conclusion.** As mentioned above, word embedding has its drawbacks. For example, to determine the lexical combinations of words, all the words of the Constitution must be in the dictionary. However, the lexical combinations of the words of these vectors can be influenced by the words before and after them. Therefore, phrase embedding should be used to increase the accuracy of the system.

Such method can help to raise the quality of research writing in any field and it can be taught as a separate discipline for improving scientific basis of the specialists. However, this process takes a lot of time to upload as many as possible texts in the corpus for further dictionary forming. This research needs more time and experiments in the future.

*References*

[1] Bondareva I.V., Lagerev D.G. Issledovanie metodov vektornogo predstavlenija tekstovoj informacii dlja reshenija zadachi analiza tonal'nosti /Informacionnye tehnologii intellektual'noj podderzhki prinjatija reshenij: Vserossijskaja nauchnaja konferencija – Ufa-Stavropol, 2018. – 10-15 p.

[2] Exploring Kotlin. Backpropagation Step by Step. [Electronic resource] URL: https://hmkcode.com/ai/backpropagation-step-by-step/ (accessed date: 10.06.2020).

[3] KDnuggets. Implementing Deep Learning Methods and Feature Engineering for Text Data: The Skip-gram Model. [Electronic resource] URL: https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-skip-gram.html (accessed date: 10.06.2020).

[4] Kusner, Matt, et al. From word embeddings to document distances. International Conference on Machine Learning. Volume 37. – Washington University, St. Louis., USA – 2015. [Electronic resource]: URL: http://proceedings.mlr.press/v37/kusnerb15.html. (Accessed date: 10.06.2020).

[5] McCormick C. Word2Vec Tutorial - The Skip-Gram Model. [Electronic resource] URL: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/. (Accessed date: 10.06.2020).

[6] Meyer D. How exactly does word2vec work? – July 31, 2016. – Pages 1-18. [Electronic resource] URL: https://pdfs.semanticscholar.org/49ed/be35390224dc0c19aefe4eb28312e70b7e79.pdf. Accessed date: 10.06.2020

[7] Mikolov, Tomas, et al. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. – Cornell University, New York, USA. – 2013. [Electronic resource]: URL: https://arxiv.org/abs/1310.4546. (Accessed date: 10.06.2020).

[8] Mohamed E.H., Shokry E.M. QSST: A Quranic Semantic Search Tool based on word embedding. Journal of King Saud University. Computer and Information Sciences. – Riyadh, Saudi Arabia. – 4 January 2020. [Electronic resource]: DOI: 10.1016/j.jksuci.2020.01.004 (Accessed date: 10.06.2020).

[9] Ould-Amer N., Mulhem Ph., Gery M., Abdulahhad K. Word Embedding for Social Book Suggestion. Clef 2016 Conference. Volume 1609. – Evora. – 09.05.2016. [Electronic resource]: URL: http://ceur-ws.org/Vol-1609/16091136.pdf (Accessed date: 10.06.2020).

[10] Wohlgenannt G., Chernyak E., Ilvovsky D. Extracting Social Networks from Literary Text with Word Embedding. Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). – Osaka, Japan. – December 11-17.2016. – pp. 18–25. [Electronic resource]: URL: https://www.aclweb.org/anthology/W16-4004.pdf (Accessed date: 10.06.2020)

**Зерттеу жұмыстарын жазудағы word embedding**

***Д.А.Аязбаев[1], А.Муратқызы[1], Р.З.Жумалиева[1]***

*[1]Сүлеймен Демирел Университеті (Қаскелең, Қазақстан)*

*Аңдатпа*

Білім берудегі ғылыми негіз ретінде саналатын зерттеу жұмыстарын жазуда тиісті термин немесе анықтаманы таңдаумен байланысты кейбір қателіктер орын алады. Бұл зерттеуде word embedding қолдану арқылы мұндай мәселені шешу және ғылыми мақалаларды жазу пәнін оқытқанда пайдалану болатындығы туралы айтылады. Word embedding – сөз көрінісінің формасы, мұнда бір сөздің векторы және оның координаттары болады. Жақын мағынасы бар сөздер ұқсас бағытқа ие бола тұра лексикалық тіркесулерді көрсетеді. Лексикалық қатынастарды есептеу үшін екі сөздің векторлары арасындағы бұрыштың косинусы қарастырылады. Тіркесулері көп сөз тіркестерінің мәні 1-ге тең. Ал лексикалық жағынан сәйкес келмейтін сөздер шамамен -1 мәні болуы керек.

Жүйені тексеру үшін Қазақстан Республикасы Конституциясының мәтіні пайдаланылды. Атап айтқанда, Конституция бабының мағынасына қатысы жоқ сөздер енгізіліп, жүйе бұл сөздерді анықтауы керек еді. Біраз сөздерді анықтағанда жүйе жоғары дәлдікті көрсетті, бірақ кейбір сөздерді төмен дәлдікпен тапты. Мұндай фактор, енгізілген сөздер мағынасы жағынан байланысты болмаса да, олар көршілерімен лексикалық тіркесуі болуы мүмкін болғандықтан көрсетілді.

Бұл зерттеу Қазақстан Республикасы Білім және ғылым министрлігінің «Отандық білім беруді модернизациялау жағдайында көптілді IT маманының құзыретті инновациялық моделін әзірлеу және енгізу» атты гранттық жобасы аясында жүзеге асырылды.

*Түйін сөздер:* сөздердің лексикалық тіркесулері, нейрондық желі, Skip-gram моделі, сөздің векторы.

**Word embedding в написании научных работ**

***Д.А.Аязбаев[1], А.Муратқызы[1], Р.З.Жумалиева[1]***
*[1]Сулейман Демирель Университет (Каскелен, Казахстан)*

*Аннотация*

При написании научных работ, что является научным фундаментом в образовании, возникают ошибки, связанные с выбором подходящего термина или определения. В этом исследовании обсуждается решение данной проблемы с помощью word embedding – форма представления слова, где слово имеет вектор и координаты. Данный метод может быть использован при преподавании дисциплины написания научных статей. Слова с близким значением имеют сходное направление, показывая лексическую совместимость. Для расчета лексических отношений учитывается косинус угла между векторами двух слов. Значение высокосовместимых словосочетаний равно 1. При этом лексически несовместимые слова должны приблизительно иметь значение -1.

Для проверки системы использовался текст Конституции Республики Казахстан. В частности, были вставлены слова, которые не имеют отношения к смыслу статьи Конституции, и система должна была идентифицировать данные слова. Для некоторых слов система показала высокую точность, с другими – низкую. Такой фактор объясняется тем, что даже если вставленные слова не имеют значения по смыслу, они могут быть лексически совместимыми с соседними словами.

Данное исследование проводится в рамках грантового проекта Министерства образования и науки Республики Казахстан «Разработка и внедрение инновационной компетентностной модели полиязычного IT-специалиста в условиях модернизации отечественного образования».

*Ключевые слова:* лексическая совместимость слов, нейронная сеть, модель Skip-gram, вектор слова.

*Поступила в редакцию 18.06.2020*

*МРНТИ 14.35.17.*

*Т.Т. ДАЛАЕВА[1], Б.С. БАЛГАЗИНА[1], С.Г. БЕЛОУС[1], Ш.М. УЛДАХАН[1]*

*[1]Казахский национальный педагогический университет имени Абая (Алматы, Казахстан),*
*tenliktd@mail.ru; bakhitgul555@mail.ru; sbelous978@gmail.com; uldahan_titan@mail.ru*
*https://doi.org/10.51889/2020-3.2077-6861.04*

## СТУДЕНЧЕСКИЙ ЦИФРОВОЙ НИР-CLUB

*Аннотация*

Современные условия удаленного общения и обучения повлекли за собой распространение цифровых технологий, когда педагоги активно вовлечены в процесс поиска новых знаний и освоения эффективных инструментов цифровизации. Отвечая современным вызовам, авторы предлагают создать на базе КазНПУ имени Абая «Студенческий цифровой НИР-club», целью которого является стимулирование научно-исследовательской и учебно-познавательной деятельности студентов, культивирование энтузиазма студентов, формирование критического мышления и лидерских качеств, навыков творческой активности, работы в команде.

В основу проекта «Студенческий цифровой НИР-club» заложена дистанционная научная деятельность и элементы геймификации, которые позволят привить обучающимся интерес к научно-исследовательской работе. Новизна проекта в том, что впервые на качественно новом уровне предпринимается попытка модер-