Journal of Engineering and Applied Sciences 14 (20): 7714-7721, 2019
ISSN: 1816-949X

Analytical Processing of Textual Resources and Documents in the Kazakh Language

A. Shormakova, Zh. Zhumanov, B. Abduali, D. Rakhimova and D. Amirova
Department of Information Technologies, Al-Farabi Kazakh National University, Almaty, Kazakhstan
shormakovaassem@gmail.com

Abstract: The theme of this research is intelligent search engines that can search and extract new information from text data in the Kazakh language and education. The significance of the research topic due to the growing amount of data represented in digital form which provide the ability to access various sources of electronic documents. The use of intelligent search engines will allow you to meet the information needs of users. In this regard, the development of information-analytical search engines that allows you to work with data in the Kazakh language is relevant. The goal of this research is to develop efficient algorithms and models for intelligent search systems, based on modern technologies in the field of information retrieval and natural language processing teaching them.

Key words: Search, learning, Kazakh text, intelligent, modern technologies, teaching

INTRODUCTION

Overview modern information retrieval: The problem of finding a document that meets certain criteria occurs in any data warehouse that contains more than one document. It is obvious that the solution of this problem is somehow confined to those which are used in the design of storage systems. You can specify two basic ways using a hierarchical model, the use of hypertext models.

The use of a hierarchical multilevel model involves the categorization of information resources. To select the path to the desired document uses the description drawn up by the support of this system. Hypertext Model allows to link documents links which are located directly in the text. These two models have obvious drawbacks. As multi-level categorization and the placement of links is performed by highly qualified specialists, the volume treated in this way documents may not be very large. For this reason, suffers the relevance of the description of the array of documents. In addition, related documents limited to any one subject area which moreover, the user of the system may be a different idea than the originator of the subject. Finally, find the correct document to the user of such systems will be required to view many documents with useful information which will only be links to other resources. These problems become particularly acute when large volumes of information, high speed of updates and the high heterogeneity of user needs. The process of finding information is the sequence of steps that lead through the system to a certain result and allowing to assess its completeness. Since, the user usually does not have comprehensive knowledge about the information content of the resource which searches, then evaluate the adequacy of a query expression as well as the completeness of the result, it can based only on external measurements or on the intermediate results and generalizations, comparing them.

The accuracy and completeness of the search depends not only on characteristics of the IRS but also on how to create a query. The ideal query can be executed by the user, fully familiar with the subject area that interests him and used IRS. To improve the quality of search, you can use a variety of methods. The most used of them is the use of Boolean operators AND, OR, NOT. Using Boolean operators a fairly easy way to increase the relevance of the documents issued but it has its drawbacks. The main one is the bad scalability. The application of the operator can greatly narrow the results and the operator greatly expanded.

The degree of accuracy and completeness of the search depends on how common terms are used in the formulation of the query. May be wrong use as the most general of terms (increases the level of informational noise) and too specific terms (reduces the completeness of the search). The use of very specific terms may lead to the fact that in the dictionary of the IRS may not be of this term.

Corresponding Author: A. Shormakova, Department of Information Technologies, Al-Farabi Kazakh National University, Almaty, Kazakhstan, shormakovaassem@gmail.com

Currently, there are enough powerful information system that more or less satisfy the information needs of users. However, the main disadvantages of most systems are the limitations of the analytical work with the resources and integration of resources within each system and with external systems (often not taken into account international standards and recommendations, low interoperability) (Anonymous, 2018a-d).

There are quite a number of algorithms for intelligent processing of text documents. Each of them has their own metric by which to measure the results of clustering. Description of algorithms of cluster analysis of texts is given in and on the website http://www. basegroup. ru/library/analysis/clusterization/datamining/ and in this study we propose their classification by dividing into two large groups:

- Algorithms flat clustering
- Hierarchical clustering algorithms

The first group includes algorithms that use the method of quadratic errors, the k-means algorithm (k-means), graph theory methods, methods based on the concept of density, neural network methods, etc., the second group includes algorithms agglomerative hierarchical clustering (divide bottom-up) clustering methods single and full connectivity, the clustering pair-group average and dividing algorithms (divide top-down) clustering using suffix trees.

In many leading scientific centers and commercial companies have active projects on creation of systems of semantic query processing which are improving and developing new protocols, technologies, programming environments, agents, languages, user interfaces, methods, distributed knowledge. For example, the DBpedia project, aimed at extracting structured information from data generated within the project Wikipedia, named one of the most successful examples of the use of technologies of the semantic processing of data Tim Berners-Lee. Almost all well-known company IBM, Adobe or Oracle, actively use the technology of the Semantic Web in their products to solve data management tasks. Microsoft invests hundreds of millions of dollars in project interactive of network resources. NET which reflects their idea of the near future internet. The system allows for automated exchange of network resources between separate programs, applications, databases, users (Buneman *et al*., 1997; Sint *et al*., 2009; Tukeyev, 2015; Wang and Guo, 2012; Myers and McGuffee, 2015).

Issues of development and creation of information retrieval systems that are able to automatically search and extract new information from semi-structured data to the scientific community, involved in various research groups. Thus, as development tools these systems use technologies such as JSP, JavaScript, PHP, MySQL database server. With the undisputed advantages of these systems with increasing volume of processed data there is a noticeable decrease in their performance (Shokin *et al*., 2010).

MATERIALS AND METHODS

Overview of the modern methods of semantic processing of textual resources: Currently, among the areas of information retrieval, a special place is the class of problems concerning smart search which involves: modeling representation of documents and queries, search and knowledge representation in digital form, classification (categorization) of texts, clustering, semantic knowledge extraction from texts.

Taking into account constant growth of digital data plays an important role improving the quality of information retrieval through the use of new semantic technologies and methods.

Big data-developed various algorithms and methods for machine solving this problem, so as to carry out the analysis manually enable the volumes of data. Any natural language in their own complicated, unique and versatile, so, extracting data from documents and textual resources is a big and time consuming job that requires preprocessing.

During the design of the module analytical processing of textual resources and documents were studied different methods and models. Such as fastText, GloVe, Word2Vec.

FastText is a library to the study of the attachment of words and text classification, laboratory of AI research at Facebook. The model is an unsupervised learning algorithm for obtaining vector representations for words. Facebook provides pre-trained models for 294 languages. This program is written in Python and C++ (Lukashevich, 2011).

A popular idea of modern machine learning is the representation of word vectors. These vectors capture the hidden information about the language such as the word analogy or semantics. It is also used to improve the performance of text classifiers. Tool fastText, you can build that dictionary vectors. fastText provides two models for calculating representations of words, skipgram and cbow ("continuous package of words").

GloVe, invented from the Global Vectors is a model for distributed representation of words. The model is an unsupervised learning algorithm for obtaining vector

representations for words. Metrics of similarity used for the estimates of nearest neighbors, creating a single scalar which quantifies the relationship between the two words. This simplicity may be problematic because the two words are almost always demonstrate more complex relationships than can be captured by a single number.

Word2Vec tool used for the analysis of the semantics of natural languages which is a technology that is based on distributional semantics and vector representation of words. This tool makes it faster than using other methods to vector on huge amounts of linguistic material.

In scientific research, Masterman (1961) described the basic ideas of information retrieval. Presented various options for finding statistics text which include counting the number of occurrences of words in documents and the frequency of adjacency of words and the new model architectures for computing continuous vector representations of words from very large data sets. Was explored as vector representations of words obtained by various models on a set of syntactic and semantic language tasks. Mikhailov *et al*. (1976) shows the use of language models neural network to the problem of calculating the semantic similarity for the Russian language. Describes the instruments, corpora and the results.

Vector representations of words trained using Word2Vec models are semantic meanings and are useful in various tasks of NLP (Natural Language Processing-Natural Language Processing). In the detailed descriptions and explanations of the equation parameters of Word2Vec models including models CBOW and skip-gram as well as advanced optimization techniques including hierarchical softmax and negative sampling. In the study presents the results of the Word2Vec algorithm for synthetic agglutinative Kazakh language. The main difficulties of the implementation of the algorithm was associated with the requirement of normalization of the text (Drakshayani and Prasad, 2013; Verma and Vuppuluri, 2015; Kenesbaev, 1977; Vinogradov, 1977; Anonymous, 2018a-f; Mikolov *et al*., 2013; Rong, 2014).

The most difficult problem arising from the creation of intelligent systems is to develop methods of automated knowledge extraction from documents in natural language. This problem is still, apparently does not have any general solution, since, the construction of this solution involves in particular, sufficiently, accurate modeling of cognitive human activity and the availability of powerful tools like syntactic and semantic analysis of texts.

On the basis of the research of the developed models applied the most for analytical processing of textual resources is the Word2Vec method. Below is represented the development of methods of the module analytical processing and implementation based on this approach.

The learning algorithm based on the method of WORD2VEC: Word2Vec includes a set of algorithms to calculate the vector representations of words, assuming that words used in similar contexts, i.e. are semantically close:

$$(wv*wc)(wc1*wv) \quad (1)$$

- In the numerator: the proximity of words of context and target words
- In the denominator the proximity of all other contexts and target words

The learning algorithm in Word2Vec there are two main learning algorithm: CBOW(Continuous Bag of Words) is a "continuous bag of words" model architecture which predicts the current word based on the surrounding context.

CBOW predicts the word from the local context:

- Inputs one-hot representations of words dimension V
- The hidden layer B matrix representations of the words W
- Output of the hidden layer is the average of vectors of words in context
- The output is a rating uj for each word and taking the softmax value which is determined by the following formula:

$$p (i|c1, ..., cn) = exp (uj) j' = 1Vexp (uj')$$

Skip-gram works differently: It uses the current word to predict the surrounding words.

Skip-gram predicts the words the context of the current word: A word predicted by the context of the central-now several multinomial distributions and softmax for each word context:

$$p (ck|i) = exp (kck) j' = 1Vexp (uj')$$

User Word2Vec has the ability to switch and choose between the algorithms. Word order context does not affect the result in any of these algorithms. The calculation uses an artificial neural network. During training, the algorithm generates the optimal vector for each word using the CBOW or skip-gram. A method of

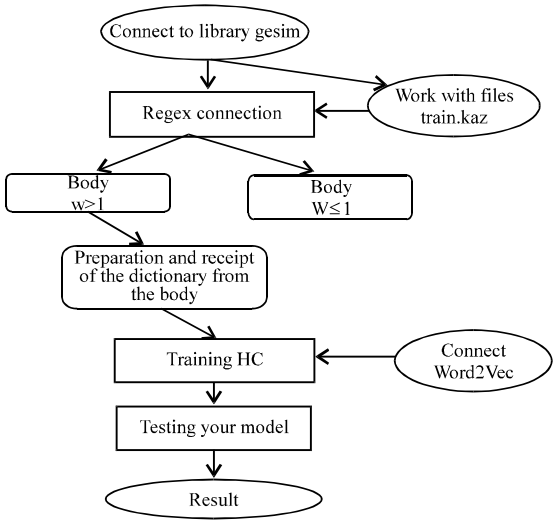

Fig. 1: The algorithm of the analytical module data processing

representing words as vectors is used for clustering words and identifying their semantic proximity, i.e., shares unrelated words and connect associated that helps in the tasks of clustering and classification of texts. Obtained at the output of the coordinate representation of the vectors of the words allow you to calculate the "semantic distance" between words. As the tool Word2Vec is based on the neural network training to achieve the most efficient operation, it is necessary to use large corpora for training. This allows to improve the quality of predictions.

In distributional semantics, words are typically represented as vectors in a multidimensional space of their contexts. Semantic similarity is computed as cosine similarity between the vectors of two words and can take values in the interval (-1, ..., 1) (in practice often used only values above 0). A value of 0 means roughly that there are no words of similar contexts and their values are not related to each other. A value of 1 in contrast, indicates the full identity of their contexts and hence, about the close value.

Recently, the interest in distributional semantics has increased significantly. This is mainly due to the new learning algorithms on large corpora: the so-called word embedding models (often their training uses simple artificial neural network). The result is a compressed vector for words that can be used for a variety of computer-linguistic tasks.

Implementation of the module analytical processing of text data: For the implementation of the module analytical processing of textual resources and documents in the Kazakh language was developed a model based on Word2Vec. The algorithm of the model shown in Fig. 1. Module for analytical processing of textual resources and documents for the Kazakh language consists of 3 stages:

- Preparation of input data
- Model training
- Work with trained model

Stage 1; Preparation of input data in turn consists of the following steps: For learning module, you need to prepare the case. The case is selected and processed according to certain rules, the set of texts used as a base for language research. They are used for statistical analysis and test statistical hypothesis, validation of linguistic rules in the language.

Search body may issue: All consumption of the selected word in the immediate context, on the basis of what the translator can choose a synonym, if you transfer or collocations. The frequency of words in a particular field of knowledge. Words that are often next to the selected word.

Was assembled by the Kazakh monolingual corpus. For training the model was trained monolingual Kazakh case which is in the file train_kaz. Detailed description of the collection of monolingual housing shown in this study.

Stage 2; Learning models: Next step in the implementation of the model is learning. To train the model, specify the following parameters:

- The dimension of the feature vectors is 100
- The maximum distance between the current predicted word in the sentence is 5
- The minimum level of learning 1
- The threshold cut off frequency of 4 words

After training save the model in model_kaz.model.

Stage 3; Work with the trained model: To work with the trained model and connection Word2Vec: model_kaz = Word2Vec.load("model_kaz.model").

In the result is a trained model you can search by the meaning of similar words. For more effective operation of the model and obtaining good results requires large monolingual case in the Kazakh language, since, Word2Vec is based on the neural network training. You can set the input as one word or several. Figure 2 and 3 shows an example of the model (Kutuzov and Andreev, 2015; Kalimoldayev *et al*., 2015; Schwa, 1975; Anonymous, 2019).

```
array([ 0.9991543 ,  0.21965139,  0.29315677,  0.18888853,  0.47598684,
       -0.28379786,  1.1473682 , -0.6190352 ,  1.0921485 , -0.3379822 ,
       -0.9026812 , -0.9696143 , -0.21033779, -1.0946484 , -0.14725323,
       -0.5960371 , -0.94785064, -0.0186517 ,  0.9273225 ,  0.02986972,
       -0.5971313 ,  1.6572342 , -0.5005268 , -0.72290874,  1.3720883 ,
        0.3576381 , -0.25446084, -0.6820295 , -0.05884275,  0.04245997,
       -1.2486485 ,  0.5666453 , -0.82413435, -0.516167  , -0.2035349 ,
       -0.65919286,  1.1125271 ,  0.79175997, -0.38865917,  0.13109162,
       -1.3794425 , -0.09773538,  1.5038078 , -0.22719735, -0.6705981 ,
       -0.01339606,  0.4934965 ,  0.36428472,  0.12966971, -0.0571641 ,
       -0.6472839 ,  0.6247625 , -0.47967348, -0.17849082,  0.06311992,
       -0.4211008 ,  0.21196847, -1.3122642 , -0.23150532,  0.47074622,
        0.63370025, -0.96714115, -0.33243644,  1.2825935 , -0.38998842,
       -0.5900854 ,  0.17189564, -0.19655013, -1.119008  , -1.2662848 ,
       -0.7881708 , -0.06033329,  0.6032156 , -0.3427293 ,  0.83609366,
       -0.6120512 , -0.12485034,  0.9940732 ,  0.58395654,  0.34109122,
       -0.35265407, -0.7005883 , -0.6270258 ,  0.5748394 ,  0.1292389 ,
        0.59908956,  0.7373656 ,  0.3227748 , -1.2803192 ,  0.6388992 ,
       -0.3058275 , -0.64198923, -0.43613726, -0.6778025 , -0.62231904,
       -1.255485  , -0.07670641,  0.7747939 ,  0.5336992 , -1.0332865 ],
      dtype=float32)
```

Fig. 2: Example of the operation of the model

```
[('1993', 0.9606807231903076),
 ('mausymda', 0.955879271030426),
 ('assambleiasynyn' , 0.955683708190918),
 ('nauryzda ', 0.9535720944404602),
 ('so'zine', 0.9529553651809692),
 ('k'azan', 0.9526992440223694),
 ('1994', 0.9519071578979492),
 ('shyrsynyn', 0.9517829418182373),
 ('mamyrynda', 0.9508242607116699),
 ('mamyrda', 0.9497315287590027)]
```

Fig. 3: Example models for the two words

Result of training for a single word, the result of training for 2 words: model_kaz.most_similar (positive = ["zhyldyn"", "bas"], topn = 10). The software part of the module written in the Python programming language of version 3. For feature models was connected to the library Gensim. Gensim-originally a library for topic modeling texts (Abiteboul *et al*., 1999). Further, for more effective work model based on Word2Vec is based on the neural network training, it is planned to increase the volume of the housing. Since, large volumes of the buildings allow you to improve search quality.

RESULTS AND DISCUSSION

Develop and implement a method of collecting synonyms Synonymy and its importance in information retrieval: One of the most important elements influencing the results of searching information is a thesaurus of keywords which involves expanding subject area due to synonymy and formation on this basis of a thesaurus of synonymy. In the 1970's and thesauri have been actively used for information retrieval tasks. In such thesauri words are mapped to descriptors through which semantic relationships are established. The first modern English thesaurus was created by Peter mark Roget (English) in 1805. It was published in 1852 and has since been used no reprints. There are also electronic dictionaries of synonyms and thesauri, English language, etc. Unfortunately, for Turkic languages electronic language resources do not exist in the public domain which would then be used in various applied problems of artificial intelligence. There is important to use the knowledge of synonymy to search for and fullness of meaning of words in the Kazakh language.

Development of a thesaurus of synonymy of the Kazakh language: The main challenge in the determination of synonyms is their determination to lexical and morphological features are automated is not possible.

Table 1: List of many meanings (synonyms)

Multi-valued words in English languages	Main1 in Kazakh language	Main 2 in Kazakh language	Main 3 in Kazakh language
String	Жоп	Jip	ishek
Order	Ret	Jarlyk'	orden
Part	Bo'lik	Partia	dene
Small	Kishkentai	U'sak'	shagyn
Thing	Zat	Na'rse	dunie
Discover	Baikau	Ashu	tabu
Information	Ak'parat	Habar	ma'limet
Field	O'ris	Dala	alan'
Present	Tanystyru	Ko'rsetu	u'synu
Observe	Bakylau	K'arau	baik'au
Make	Jasau	K'u'ru	isteu
Go	Baru	Ketu	juru
So	Solai	Osylai	bu'lai
Call	Atau	Shak'yru	k'on'yrau shalu
Set	K'oiu	Otyrg'yzu	ornatu

Table 2: Translations and the frequency of words in the context

Variables	Frequency	Values
f_1	amb_word_1	$freq\_of\_f_1$
f_2	amb_word_1	$freq\_of\_f_2$
f_3	amb_word_1	$freq\_of\_f_3$
…	…	…
f_n	amb_word_n	$freq\_of\_f_n$

Synonyms should be context-dependent that is the relationship of synonymy is possible between words. In this study, the researchers developed a hybrid approach based on the method of maximum entropy in the practical implementation of the semantic cube.

To solve this problem were used the linguistic resources of the English language which made the implementation of the approach. Below is an example of multi-valued words in English that have different meanings in Kazakh language (Table 1).

This list is based on a parallel English-Kazakh corpus that identifies a particular translation of the words. Found synonyms (multi-valued words) and their adjacent words are filled in the table. Table 2 of possible translations in the contest can be summarized as follows:

- Here, amb_word-multivalued words from the initial context, t transfers
- f-context of the target language, freq_of_f
- Frequency of occurrence of ambiguous words in context

Using a table of frequencies while testing for the right words, it calculates the probability and selects the corresponding translation. The formula which determines the translation according to the context:

$$Ps\left(tVc\right)=\frac{c(fi)}{\sum_{i=1}^{r} fi}$$

Where:

(c (fi)) = The frequency of context

$\sum_{i=1}^{r} fi$ = Amount needed possible factor contexts

az zhetkilikti shamaly shekteuli

karikatura maktan tus nazar audar ul'gait men

sypaiy meii'r zhaksy a'dilet

buryn k'ara de k'anshalyk'ty kel

densaulyk' paida a'l-auk'at k'yzyk' progress

molshylyk' o'rkende bailyk' k'az

…|

Fig. 4: Automated directory Kazakh synonyms

Definition of synonyms with their equivalents, constructed a table of frequencies. For the choice of words we use Eq. 2 and 3:

$$\hat{t}=\begin{cases} t_1 = prob\_f_{11}+prob\_f_{12}+prob\_f_{13}+,...,+prob\_f_{1n} \\ t_2 = prob\_f_{21}+prob\_f_{22}+prob\_f_{23}+,...,+prob\_f_{2n} \\ ... \\ t_n = prob\_f_{n1}+prob\_f_{n2}+prob\_f_{n3}+,...,+prob\_f_{nn} \end{cases} \quad (2)$$

The biggest argument finding probabilities of synonyms is determined by the function:

$$\text{argmax. } t = \text{argmax } P(t1,t2, ..., tn) \quad (3)$$

This defines various values that are written in the semantic dimension of the cube. The cube is directly dependent on the size of the English-Kazakh parallel corpus. The semantic form of a multidimensional cube is presented in Fig. 4. The growth of the size of the case affect the quality of the right synonym of the words that is more than a catalogue of words of synonyms the more accurate and better will find and identify the right synonyms. With the implementation of this approach has been used parallel English Bkazakh corpus and open source online dictionary of English synonyms (Thesaurus.com). Developed a thesaurus of synonymy, based on the algorithm of maximum entropy and in the practical implementation of the method of semantic Cuba.

Using this method was automated system of collection of synonyms and similar in meaning of meaning of words. Was added to the database up to 9000 entries of synonyms of the Kazakh language.

CONCLUSION

The results obtained on the development of methods and models of the module analytical processing of textual resources and documents in the Kazakh language:

- Researched different types of methods and models for semantic analysis used for operation of the module analytical processing of textual resources and documents
- Collection monolingual corpora of the Kazakh language to train the Word2Vec model
- Implemented pre-processing text data for use as input data, the selection of words that are similar in vector representation
- The module for analytical processing of textual resources and documents in the Kazakh language is implemented based on the model Word2Vec
- Developed thesaurus of synonyms of the Kazakh language of module analytical word processing
- Developed automation replenishment system database of synonyms of the Kazakh language

The experiments showed good values for the learning system analytical processing of data. In the future, increase quality due to the increase in the input and casing of the Kazakh language and through the creation of a marked body of the Kazakh language.

ACKNOWLEDGEMENT

This research performed and financed by the grant Project IRN AP05132950 “Development of an information-analytical search system of data in the Kazakh language”, awarded to the Republican State Enterprise (RGP) on the right of economic management (PVC) «Institute of Information and Computational Technologies».

REFERENCES

Abiteboul, S., P. Buneman and D. Suciu, 1999. Data on the Web: From Relations to Semistructured Data and XML. 1st Edn., Morgan Kaufmann Publishers, Burlington, Massachusetts, USA., ISBN: 9781558606227, Pages: 258.

Anonymous, 2018. Algorithm Word2Vec. Megaindeks, Putilkovo, Russia. https://ru.megaindex.com/ support /faq/word2Vec.

Anonymous, 2018a. Basic snowball stemming algorith for kazakh language. GitHub Inc., San Francisco, California, USA., https:// github.com/ iborodikhin/stemmer-kaz.

Anonymous, 2018b. Battles of Lexington and concord. Wki Media project, San Francisco, California, USA.

Anonymous, 2018c. The principle of maximum entropy. Wiki Media, San Francisco, California, USA.

Anonymous, 2018d. Why in search without linguistics cannot do?. Transcendental Meditation, Russia. https://habr.com/en/company/yandex/blog/224579/

Anonymous, 2018e. Word2Vec. Wikimedia Foundation, San Francisco, California, USA.

Anonymous, 2019f. What is fastText?. Facebook, Inc., Menlo Park, California, USA. https://fasttext.cc/

Buneman, P., S. Davidson, M. Fernandez and D. Suciu, 1997. Adding structure to unstructured data. Proceedings of the International Conference on Database Theory, January 8-10, 1997, Springer, Berlin, Germany, ISBN:978-3-540-62222-2, pp: 336-350.

Drakshayani, B. and E.V. Prasad, 2013. Semantic based model for text document clustering with idioms. Intl. J. Data Eng., 4: 1-13.

Kalimoldayev, M.N., K.C. Koibagarov, A.A. Pak and A.S. Zharmagambetov, 2015. The application of the connectionist method of semantic similarity for Kazakh language. Proceedings of the 12th International Conference on Electronics Computer and Computation (ICECCO), September 27-30, 2015, IEEE, Almaty, Kazakhstan, ISBN: 978-1-5090-0199-6, pp: 1-3.

Kenesbaev, S.K., 1977. Phraseological Dictionary of Kazakh Language. Nauka Publisher, Russian, Pages: 711.

Kutuzov, A. and I. Andreev, 2015. Texts in, meaning out: Neural language models in semantic similarity task for Russian. Comput. Lang., 1: 1-12.

Lukashevich, N.V., 2011. Thesauruses in Information Retrieval Tasks. Moscow State University Publishing House, Moscow, Russia, ISBN:9785211059269, Pages: 508.

Masterman, M., 1961. Semantic message detection for machine translation, using an interlingua. Proceedings of the International Conference on Machine Translation of Languages and Applied Language Analysis, September 5-8, 1961, National Physical Laboratory, Teddington, England, pp: 438-474.

Mikhailov, A.I., A.I. Chernyi and R.S. Gilyarevskyi, 1976. Scientific Communications and Informatics. Nauka Publishers, Moscow, Russia, Pages: 435.

Mikolov, T., K. Chen, G. Corrado and J. Dean, 2013. Efficient estimation of word representations in vector space. J. English Lit., 1: 1-12.

Myers, D. and J.W. McGuffee, 2015. Choosing scrapy. J. Comput. Sci. Colleges, 31: 83-89.

Rong, X., 2014. Word2vec parameter learning explained. Comput. Lang., 1: 1-21.

Schwa, B., 1975. Kazakh Ton Synomer SZDG. Mektep Knizhnyy Magazin Almaty, Kazakhstan, Pages: 236.

Shokin, Y.I., A.M. Fedotov, V.B. Barakhnin and O.L. Zhizhimov, 2010. Problems Finding Information. Nauka Publisher, Russia, ISBN:9785020189690, Pages: 197.

Sint, R., S. Schaffert, S. Stroka and R. Ferstl, 2009. Combining unstructured, fully structured and semi-structured information in semantic wikis. Proceedings of the 4th and 6th Joint Conference on Semantic Wiki (SemWiki 2009) and European Semantic Web (ESWC 2009), June 1, 2009, Morgan Kaufmann Publishers, Hersonissos, Greece, pp: 1-15.

Tukeyev, U., 2015. Automaton models of the morphology analysis and the completeness of the endings of the Kazakh language. Proceedings of the International Conference on Turkic Languages Processing (Turklang-2015), September 17-19, 2015, Kazan, Tatarstan, pp: 91-100.

Verma, R. and V. Vuppuluri, 2015. A new approach for idiom identification using meanings and the Web. Proceedings of the International Conference on Recent Advances in Natural Language Processing, September 5-7, 2015, Hisarya, Bulgaria, pp: 681-687.

Vinogradov, V.V., 1977. The Main Types of Phraseological Units in the Russian Language. Lexicology and Lexicography, Russia, Pages: 135.

Wang, J. and Y. Guo, 2012. Scrapy-based crawling and user-behavior characteristics analysis on Taobao. Proceedings of the 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, October 10-12, 2012, IEEE, Sanya, China, ISBN:978-1-4673-2624-7, pp: 44-52.