Hybrid approach for the semantic analysis of texts in the Kazakh language

Diana Rakhimova[0000-0003-1427-198X], Asem Turarbek[0000-0002-4793-0446]
and Leila Kopbosyn [0000-0001-6224-2193]

Al-Farabi Kazakh National University, Almaty, Kazakhstan
di.diva@mail.ru, turarbek_asem@mail.ru, leila_s@list.ru

Abstract. In this paper authors propose a hybrid approach for semantic analysis of text resources and documents in the Kazakh language. An overview and difficulties of analysis for the Kazakh language are presented. The developed approach consists of two main parts. The first definition of keywords (phrases) from the text, and the second, based on the data obtained, will build an annotated summarization of the text. To implement the first part of the approach, the TF-IDF algorithm was applied to extract keywords and phrases from texts. The cosine similarity of the sentence data in the Kazakh language was calculated to determine the similarity. With the help of certain similarities semantic links in the text are determined. On the basis of the data obtained, the second part is performed - the abstraction of texts. The number of annotations directly depends on the size of the document. The linguistic corpus of the Kazakh language was collected for carrying out experiments and calculations. A study of various approaches and a hybrid approach for the semantic analysis of the Kazakh language was carried out. The practical part was implemented in Python. The article presents the results of experimental calculations.

Keywords: Kazakh language, semantic analysis, keywords, summarization.

1 Introduction

The Kazakh language belongs to the Turkic group of languages and the agglutinative class of languages, it has a complex morphological structure and a rich semantic vocabulary. Unfortunately, at the moment, the Kazakh language is a low-resource language, which hinders the development and conduct of scientific research. For the Kazakh language, the problem of semantic analysis and identification of data or facts is relevant. There are no universal approaches and methods that allow for high-quality semantic analysis, to identify data and facts from texts, etc.

Computer semantic analysis is closely related to the problem of text understanding by a machine. There are many interpretations of the concept "meaning of the text" and the task of understanding it. For example, according to D.A. Pospelov [1], the system understands the text entered into it if, from the point of view of a person (or a group of experts), it correctly answers questions related to the information contained in the text.

2 Related works

There are various scientific approaches and methods for solving the problem of semantic analysis for a particular language. Some of them will be presented below. Of course, no software can replace the analysis that a human can think of. However, the programs that are currently being developed can reduce the time spent on studying large databases. In this regard, the work of the following programs for solving problems of semantic text analysis is considered. Software offered by various manufacturers, such as Semantic LLC, Tomita-parser (Yandex), Semantic Analyst JHON, SummarizeBot API, TextAnalyst 2.0, Galaktika-ZOOM, NLP ISA Natasha »Etc. is used in different subject areas and for different languages [2-9].

For example, "Semantic LLC" is a program for editing unstructured text. The semiconductor line is graphically oriented, each node is a semantic element, and the walls represent the elements of the elements. Each attribute of a node is of great importance, the set of attributes depends on the type of element.

Tomita Parser (Yandex) is a program that allows you to extract facts from structured text. Separation of facts is based on context-independent grammar rules. And the program requires a dictionary of keywords. The parser will write its own grammar.

SummarizeBot API - The web service offers a RESTful API to handle all text and image processing tasks. It uses over 100 languages including Russian, English, Chinese, Japanese, and uses machine learning technology. The current version uses the following parameters: 1) automatically link to text; 2) Selection of keywords and conceptual documents; 3) Analysis of a sample of documents and selection of material objects and attributes; 4) Automatically detect the language of the document; 5) Obtaining unpublished data: the main text of articles, forums, forums, etc .; 6) Image processing: identification and recognition of objects in images.

"TextAnalyst 2.0" - a program developed by the research and production innovation center MicroSystems as a tool for text analysis. Text links allow you to create a semantic web of comments, expressed in processed text. The request has the ability to semantic search for fragments of text taking into account the semantic links hidden in the text. Allows you to parse text by constructing a hierarchical tree / heading topics containing text.

The scientific works [10-14] describe the basic ideas of using semantic analysis in information retrieval systems. Various options for finding text statistics are presented, which include counting the number of occurrences of words in documents and the frequency of word contiguity, and new model architectures for computing continuous vector representations of words from very large datasets. The quality of vector representations of words obtained by various models was studied using a set of syntactic and semantic language problems. In [15], the application of language models of a neural network to the problem of calculating semantic similarity for the Russian language is shown. The tools and bodies used and the results achieved are described.

The above presented software products are designed for many resource languages such as English, Spanish. Russian, etc. Unfortunately, for the Turkic languages (Kazakh, Kyrgyz, Turkish, Uzbek, etc.) there is currently no software implementation in the open access. The disadvantage of the developed systems is that they cannot be

applied to the Turkic languages, since they are agglutinative with complex morphological and lexical forms, and semantics dependent sentence structure.

The analysis of a huge amount of data can be simplified if we have keywords or keyphrases that can provide us with the basic characteristics, concept, etc. of a document. The relevant keywords and keyphrases can serve as a summary of the document and help us easily organize documents and extract them based on their contents [16]. It is necessary to distinguish two main approaches to solving the problem of automating the selection of keywords and keyphrases: the assignment of keywords and keyphrases and their extraction [17-18]. The main difference is that the first approach allows to select only those keywords and keyphrases that are contained in some provided dictionary, and the second approach involves the selection of key information directly from the text.

Keywords can be assigned manually or automatically, but the first approach is very time-consuming and expensive. Thus, there is a need for an automated process that extracts keywords from documents. There are ready-made software solutions to this problem for common languages (English, Russian, Spanish, etc.), and for the Kazakh language there are only a few and they are not in open access.

Below are some approaches and works for carrying out summarization for different languages:

The most common is the superficial approach, which takes into account title words and cue-words (ie, "important", "best" etc.) To extract response results [19].

The paper [20] presents automatic free text processing using material extraction using agent verification. For data processing, the Kmeans algorithm was used as a basis.

There is a common summarization approach based on the structural removal of parts from the text corpus. For example, the WordNet system [21].

The paper [22] presents the Cohesive Approaches, which define and consider the cohesive relationships between concepts within the text. These include synonyms, antonyms, lexical data of the language, etc.

It should be noted that at the moment one of the most popular methods of summation is graphical approaches. Two methods can be attributed to this type: LexRank [23] and TextRank [24].

In [25], the graph approach of summarization a text document is also presented. The difference between this approach is that it simultaneously takes into account local coherence, importance, and redundancy.

The next type of approach is based on machine learning. With this approach, the resulting document results can be transformed into a controlled or semi-controlled learning task. This method requires big data to conduct training.

In the article [26], a new Seq2Seq model is presented for abstract and extractive generalization. A comparative analysis of existing approaches is carried out and it is shown that RNNs and other Seq2Seq models represent a good practical result. The main difference of this approach is at the first-time step during encoding the sequence of adding contextual information using the agent.

3 A semantic analysis based an algorithm for extracting annotation and keywords

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

To work with big data, various algorithms and methods are being developed for the machine solution of this problem, since the amount of data does not allow for manual analysis. Any natural-language is complex, unique, and multifaceted in its own way, therefore, extracting data from documents and text resources is a large and time-consuming work that requires preliminary processing.

This part will present a hybrid approach to the semantic analysis of text resources and documents in the Kazakh language. The developed approach consists of two main parts. The first definition of keywords (phrases) from the text, and the second, based on the data obtained, will build an annotated generalization of the text.

The developed hybrid approach of semantic analysis of the text in the Kazakh language consists of two main stages:

- identify keywords and phrases in the text;
- making semantic annotation of the text based on keywords.

For the first stage, it is necessary to prepare the text. To do this, lemmatization and marking by morphological properties are performed on the texts. The main task of the keyword detection algorithms is the task of finding suitable candidates, identifying attributes and ranking [29].

To rank and determine the frequency, the TF-IDF (Term Frequency - Inverse Document Frequency) indicator was used [28]. With TF-IDF, you can determine the weights for each word relative to the entire document. The words with the highest scores and are the main keywords of the text.

TF-IDF was calculated using the formula below

$$TF * IDF = TF(t, D) * IDF(t) = \frac{n_{t,D}}{\sum_k n_{k,D}} * log\left(\frac{|TS|}{|\{d:t \epsilon d\}|}\right) \quad (1)$$

where $n_{t,D}$ is the number of occurrences of the word t in the target collection D, $\sum_k n_{k,D}$ is the sum of the occurrences of all words in the target collection D, $|TS|$ is the number of documents in all used collections, $|\{d: t \in d\}|$ is the number of all documents that include the word t at least once.

According to this formula, the weight of the word is calculated. The higher the weight of a word, the higher its relative frequency of use in the collection of text. Based on this algorithm for determining keywords and properties and linguistic resources of the Kazakh language, a modified algorithm for extracting keywords and phrases was developed [13].

To find the similarity of the elements (sentences) of the text and the evaluation, the cosine similarity was applied. To calculate the cosine similarity between sentences, you need to perform the following steps: first, you need to identify all the individual words. Then the identification of the frequency of occurrence of these words in

sentences is formed and is defined as a vector. That is, the sentence itself will be represented as a set of vectors. Next, the cosine similarity function is applied to these vectors, and the cosine of the angle between the vectors is subtracted. [14, 15]

x and y are sentence vectors. Their scalar product and the cosine of the angle θ between them are related by the following relation

$$< x, y > = ||x|||y||\cos(\theta) \qquad (2)$$

Accordingly, the cosine distance is defined as

$$\rho_{cos}(x, y) = \arccos\left(\frac{<x,y>}{||x|||y||}\right) = \arccos\left(\frac{\sum_{i=1}^{d} x_i y_i}{\left(\sum_{i=1}^{d} x_i^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^{d} y_i^2\right)^{\frac{1}{2}}}\right) \qquad (3)$$

Based on the data obtained from formula 3, a matrix of the similarity values of the sentences is constructed. Next, all the offers are ranked according to the similarity matrix. The sentences with the highest weight, which are defined by keywords or phrases, will form the annotation of the document.

This proposed approach takes into account the grammatical properties and rules of the Kazakh language. The next section presents the practical results of the developed hybrid approach to semantic analysis.

4 Application of approaches and experimental results

At the first stage, 2 tasks are solved: preliminary word processing; and the division of the text into separate words and keyphrases.

The first task is language-dependent, therefore, the Kazakh language morphological feature is taken into account here. To solve this problem, a system of complete endings of the Kazakh language is used (through the morphological analyzer of the Kazakh language developed on the platform Apertium [30], we perform markup of the document), the algorithm for stemming and lemmatization for the Kazakh language [31] (implemented in the Python3 programming language). Then, a simple approach was used - the tokenization procedure, which helps to divide the whole text into separate words.

The developed algorithms and approach for hybrid semantic analysis are implemented using the Python programming language and NLTK libraries. To test the program, we have prepared a marked corpus, which consists of more than 120 text documents of various sizes and topics. First, keywords and phrases with the Tf-idf metric were defined for each text . Table 1 below shows an example of the keywords found for texts in the Kazakh language.

Table 1. Experimental data of the obtained keywords from texts in the Kazakh language.

Keywords and keyphrases	*Tf-idf metric*
Document: arabazathistory.txt, Number of words in the text: 1876	
Ливан (Lebanon)	0.03753761448295349
Көтеріліс (revolution)	0.014962316253101847
Француз (French)	0.014881951295324384
Франция (France)	0.011384757884540301
француз үкімет (the French government)	0.013168923967413456
1920 жыл (1920 year)	0.008728017814309411
келісім шарт (agreement)	0.00827156782972209
Document: okushi.txt, Number of words in the text: 3450	
Сабақ (lesson)	0.010324737893214916
Физика (physics)	0.006381991500464335
Ауылшаруашылық (agriculture)	0.003477428443091718
Мұғалім (teacher)	0.003398202016653529
сынып физика (class physics)	0.0037965546730691483
…	…

Fig. 1. An example of the operation of the algorithm for determining keywords and phrases (the measure TF and IDF are shown separately).

Fig. 2**.** An example of the operation of the algorithm for determining keywords and phrases (the measure TF-IDF is shown).

Table 2 presents the practical results of the developed algorithm for determining keywords and phrases in Kazakh texts.

Table 2. Experimental results of the developed algorithm for determining keywords for the Kazakh language

Document's name	Document volume (number of sentences)	Borderline coefficient keywords	Number of keywords	Accuracy finding
Sport.txt	87	3-8	8	84,31%
books.txt	79	3-8	8	84%
almaty.txt	96	3-8	8	79,5%
Psychology.txt	298	9-12	10	93,4%
2018biznesmen.txt	320	12-15	12	63,43%
computersciense.txt	415	15-17	12	95,03%
geoinformatika.txt	885	15-17	13	98,3%

Taking into account the limiting coefficient of determining keywords by the volume of the text, the keywords and phrases are selected according to the meaning correctly and has a not bad indicator of accuracy.

To test the operation of the developed algorithm for extracting keywords in the Kazakh language, practical experiments were conducted. In practice, two approaches were compared: the first simple summarization, the second summarization with keywords and phrases. In the experiment, more than 120 documents in the Kazakh language with various topics and volumes were processed. The time spent on identifying the text annotation directly depended on the volume of the input text. The resulting annotations are shown in table 3.

Table 3. Examples of the work of summarization approaches for texts in the Kazakh language.

Document: computer.txt		*Translate*
Summarization based on keywords	Компьютер (ағылшынша: computer — «есептегіш»), ЭЕМ (электрондық есептеуіш машина) — есептеулерді жүргізуге, және ақпаратты алдын ала белгіленген алгоритм бойынша қабылдау, қайта өңдеу, сақтау және нәтиже шығару үшін арналған машина. Компьютер шеше алмайтын есептерді ағылшын математигі Аланом Тьюринг сипаттаған болатын. Бұл ерекшелікті алғаш рет 1965 жылы «Intel» компаниясының басшыларының бірі Гордон Е Мур сипаттаған болатын. Көптеген ғалымдар компьютерді адамға ыңғайлы ондық санау жүйесінде жасап шығаруға тырысты	Computer (English: computer - "counter"), computer (electronic computer) - a machine designed to perform calculations, and to receive, process, store and output information according to a predetermined algorithm. Problems that a computer cannot solve were described by the English mathematician Alan Turing. This feature was first described in 1965 by Gordon E. Moore, one of the leaders of Intel. Many scientists have tried to build a computer in a human-friendly decimal number system
Simple summarization	Компьютер (ағылшынша: computer — «есептегіш»), ЭЕМ (электрондық есептеуіш машина) — есептеулерді жүргізуге, және ақпаратты алдын ала белгіленген алгоритм бойынша қабылдау, қайта өңдеу, сақтау және нәтиже шығару үшін арналған машина. Компьютер тек қана бағдарламада көрсетілген сызықтар мен түстерді енгізу-шығару құрылғыларының көмегімен механикалық түрде көрсетеді. 1946 жылы бұл сөздікте цифрлық компьютер, аналогтық есептеуіш машинасы және электронды компьютер түсініктерінің мағынасы ажыратылып көрсетілді. Бұл ерекшелікті алғаш рет 1965 жылы «Intel» компаниясының басшыларының бірі Гордон Е Мур сипаттаған болатын. Компьютерлер көлеміні кішірею процессі де осындай	Computer (English: computer - "counter"), computer (electronic computer) - a machine designed to perform calculations, and to receive, process, store and output information according to a predetermined algorithm. The computer displays the lines and colors shown in the program only mechanically with the help of I / O devices. In 1946, the dictionary differentiated between the concepts of digital computer, analog computer and electronic computer. This feature was first described in 1965 by one of the leaders of Intel, Gordon E. Moore. The process of reducing the size of

	жылдамдықпен жүріп келеді. Алғашқы электрондық есептеуіш машиналар көптеген тонна салмағы бар. Егер цифрлық компьютерлер дискретті сандық және таңбалық айнымалылармен жұмыс жасайтын болса, аналогтық компьютерлер келіп түсетін мәліметтер ағынын үзіліссіз өңдеуге арналған.	computers is going at the same speed. The first electronic computers weighed many tons. If digital computers work with discrete numeric and symbolic variables, analog computers are designed for continuous processing of incoming data streams.
Document: moon.txt		Translate
Summarization based on keywords	Біздің планетамызда жоқ заттардың орнын алмастыру қажет. Сол себепті адамдар Айға көз жүгіртеді. Ай топырағынан оттегі ал технологиясы жердегі зертханаларда пайдаланылған. Айдағы энергетиканы дамытудың басты бағыты. Адамдардың Айды игеруі – бұл жүзеге асыратын іс екенін көрсетті	We need to replace things that do not exist on our planet. That is why people look at the moon. Oxygen from lunar cancer and technology have been used in terrestrial laboratories. The main direction of lunar energy development. The fact that people have mastered the Moon has shown that it is a work in progress
Simple summarization	Геостационарлыорбита дегеніміз - бұл Жерден шамамен 35800 км биіктіктегі шеңберлер экваторлы орбита. Айдағы энергетиканы дамытудың басты бағыты, бұл күн энергиясын электр энергиясына өзгерту. Луноход -1» аппараты рентген телескопымен жабдықталған еді, ол арқылы галактика аралық рентген сәулелерінің ұзындықтары өлшенді. Адамдардың Айды игеруі – бұл жүзеге асыратын іс екенін көрсетті. Жердің экологиясын тазалау. Айдан әкелінген тас-топырақты зерттеу барысында, онда жер бетінде сирек кездесетін металдардың, пироксеннің, ильменитің т.б. Жерді аса зиянды қалдықтардан тазарту проблемасын шешу жолында, осы жұмыста көрсетілген бағыт, көңіл аударатындай ерекше болып отыр	A geostationary orbit is an equatorial orbit with circles at an altitude of about 35,800 km above the Earth. The main direction of lunar energy development is the conversion of solar energy into electricity. Lunokhod-1 was equipped with an X-ray telescope, through which the lengths of intergalactic X-rays were measured. The fact that people have mastered the Moon has shown that it is a work in progress. Cleaning the earth's ecology. During the study of rocks and soils brought from the moon, they found rare metals, pyroxene, ilmenite, etc. In addressing the problem of land degradation, the direction outlined in this paper is particularly noteworthy.

The table 3 shows examples of text processing using two summarization methods. From the results obtained, it can be seen that the received annotations convey the semantic concept of the text. In experiments on texts with a small volume, there were cases when the results of the two approaches were very approximate.

Figure 3 below shows the interface of the software solution for defining text annotations. The upper yellow window shows the original text in Kazakh. The total number of words and sentences are also indicated. Further down in the yellow window, you will

see the specific keywords and phrases that will be used in the text. The left blue window shows the result of the simple summarization, and the right blue window shows the result of the summarization based on keywords.

Fig. 3**.** An example of the program for determining summarization (two approaches) for the Kazakh language

Figure 4 shows the percentage of the results of the two summarization approaches. The horizontal values show the number of words in the document. And vertically, the percentage of the accuracy of determining the annotations of these texts. The analysis and accuracy of the results were carried out manually by three experts (a specialist linguist of the Kazakh language). Then the average value of the experts ' assessments was calculated.

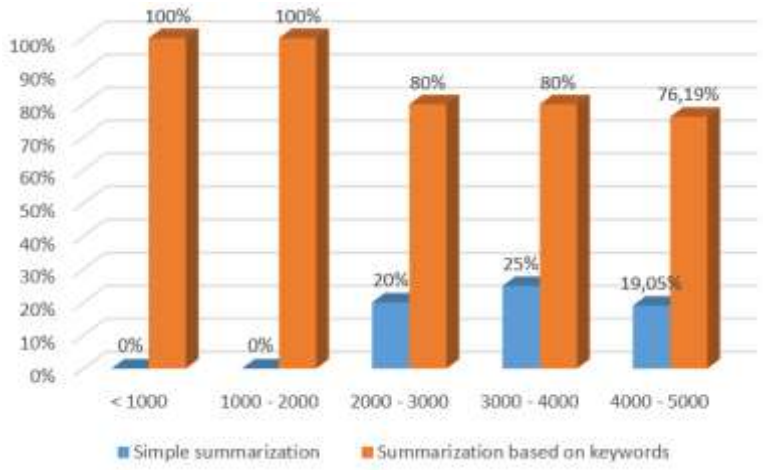

Fig. 4. The percentage of the results of the two summarization approaches.

The best result for defining the annotation of full-text documents is given by the keyword-based summarization approach. This is because keywords are used to cover sentences that have some meaning to the text, rather than simple introductory sentences. The above-developed algorithms and the method of the module are interconnected and provide an integrated approach for processing and analysis of big data in the Kazakh language.

5 Conclusion and future work

According to the results of scientific research work, the following results were obtained:

Methods and modern approaches to semantic analysis and abstraction of texts are investigated. Taking into account the peculiarity of the grammar of the Kazakh language, a hybrid semantic analysis of full-text documents was developed. This approach is based on the definition of keywords\phrases and the construction of the text annotation. The practical results of the text analysis show that this approach reveals the contextual meaning of the text. This approach can also be applied to other low-resource Turkic languages. Because it does not require large data for processing.

In the future, it is planned to use this approach in the implementation of machine translation and post-editing systems for Kazakh language.

6 Acknowledgments

This research performed and financed by the grant Project IRN AP08052421 MES RK , Project title: «Research and development of the post-editing system of the Kazakh language in machine translation», by «Research Institute of Mathematics and Mechanics» Al-Farabi Kazakh National University.

References

1. Pospelov D.A. :Ten hotspots in research on artificial intelligence Intelligent systems (MSU). (resource language – Russian). - 1996. - Vol. 1, No 1-4. - P. 47–56. (1996)
2. Semantic: http://semantick.ru/: last accessed 14.07.2020.
3. Tomita parser: http://api.yandex.ru/tomita/: last accessed 07/14/2020.
4. In the foothills of semantics: http://dworq.com/: 05/29/2020.5.AI Data Analysis Technologies for Business // https://www.summarizebot.com/summarization_business.html: last accessed 27.05.2020.
5. TextAnalyst ver. 2.0 - Program for personal text analysis: http://offext.ru/library/data/datakeeping/51.aspx: last accessed 19.04.2020.

6. Galaktika-Zoom - analytical system for respectable clients: https://www.it-week.ru/themes/detail.php?ID=52215: last accessed 16.06.2020.
7. Best Out-Of-The-Box Sentiment Analysis Tools; https://monkeylearn.com/blog/sentiment-analysis-tools/ last accessed 2020/07/25.
8. Automatic text analysis technologies (resource language – Russian). : http://nlp.isa.ru/: last accessed 26.04.2020.
9. GitHub natasha : https://github.com/natasha: last accessed 26.04.2020
10. Sonawane S.S., Kulkarni P.A.: Graph based Representation and Analysis of Text Document: A Survey of Techniques International Journal of Computer Applications. – 2014. – Vol. 96, issue 19. – P. 1-8. (2014)
11. I. Cicekli, T. Korkmaz.: Generation of Simple Turkish Sentences with Systemic-Functional Grammar. DOI: 10.3115/1603899.1603928.
12. Manning Ch.D., Raghavan P., Schütze H.: Introduction to Information Retrieval. – Cambridge University Press, NY, USA, 2008. – 210 p. (2008)
13. Efficient Estimation of Word Representations in Vector Space https://arxiv.org/pdf/1301.3781.pdf: last accessed 10.07.2020
14. Word2vec Parameter Learning Explained https://arxiv.org/pdf/1411.2738.pdf: last accessed 10.07.2018.
15. Texts in, Meaning out: neural language Models in semantic similarity tasks for russian // https://arxiv.org/ftp/arxiv/papers/1504/1504.08183.pdf: last accessed 20.04.2020.
16. Sheremeteva S.O., Osminin P.G.: Methods and models for automatic keyword extraction (resource language – Russian). Bulletin of the South Ural State University, № 1, T. 12, pp. 76-81 (2015).
17. Effective Approaches for Extraction of Keywords, http://www.ijcsi.org/papers/7-6-144-148.pdf, last accessed 2019/07/25.
18. Keyword extraction a review of methods and approaches, http://langnet.uniri.hr/papers/beliga/Beliga_KeywordExtraction_a_review_of_methods_and_approaches.pdf, last accessed 2019/07/05.
19. V. Nastase. 2008. Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 763–772. (2008)
20. García-hernández R.A., Montiel R., Ledeneva Y., Rendón E., Gelbukh A., Cruz R. 2008. Text Summarization by Sentence Extraction Using Unsupervised Learning. In: Gelbukh A., Morales E.F. (eds) MICAI 2008: Advances in Artificial Intelligence. MICAI 2008. Lecture Notes in Computer Science, vol 5317. Springer, Berlin, Heidelberg. pp. 133-143. (2008)
21. G. A. Miller. 1995. Wordnet: A lexical database for english. Commun. ACM, 38(11), pp. 39–41.(1995)
22. R. Barzilay and M. Elhadad. 1999. Using lexical chains for text summarization. Advances in automatic text summarization. pp. 111–121. (1999)
23. G. Erkan and D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22, pp. 457–479. (2004)
24. R. Mihalcea and P. Tarau. 2004. Textrank: Bringing order into texts. Association for Computational Linguistics.(2004)
25. Parveen, D., Strube, M. 2015. Integrating importance, non-redundancy and coherence in graph-based extractive summarization. In: proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), pp. 1298-1304.(2015)
26. Khatri, C., Singh, G., Parikh, N. 2018. Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent NeuralNetworks. URL http://arxiv.org/abs/1807.08000. (2018)

27. Zeng, B., Xu, R., Yang, h., Gan, Z., Zhou, W. 2020. Comprehensive Document Summarization with Refined Self-Matching Mechanism. Appl. Sci., 10, 1864. doi:10.3390/app10051864.
28. TF-IDF , https://en.wikipedia.org/wiki/Tf%E2%80%93idf last accessed 2020/07/15.
29. Hanumanthappa M., Narayana Swamy M., Jyothi N.M.: Automatic Keyword Extraction from Dravidian Language. International Journal of Innovative Science, Engineering & Technology, vol. 1, issue 8, pp. 87-92 (2014).
30. Rakhimova, D., Turganbayeva, A. Auto-abstracting of texts in the Kazakh language // Proceedings of the 6th International Conference on Engineering & MIS. – 2020. – P. 1-5 // https://doi.org/10.1145/3410352.3410832
31. Rakhimova D., Shormakova A.: Problems of semantics of words of the Kazakh Language in the information retrieval // Proceedings of the 11-th International Conference, ICCCI 2019, Springer. 2019. Part II. – P. 70-81 (2019)