BMBB

IEEE

IEEE Turkey Section

UBMK'22

Bildiriler Kitabı
Proceedings

7. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

7th International Conference on Computer Science and Engineering

14-15-16 Eylül (September) 2022 Diyarbakır - Türkiye

The Task of Synthesizing the Kazakh Language Based on the seq2seq Approach for a Question-Answer System

Diana Rakhimova[1,2]
[1]*Al-Farabi Kazakh National University,*
[2]*Institute of information and computational technologies*,
Almaty, Kazakhstan
di.diva@mail.ru

Yerkin Suleimenov[2]
[2]*Institute of information and computational technologies*,
Almaty, Kazakhstan
erken.suleimenov@gmail.com

Gulstan Akhmet[1]
[1]*Al-Farabi Kazakh National University*,
Almaty, Kazakhstan
gulstan.akhmet@gmail.com

***Abstract*—Currently, many texts are synthesized using a computer. Sentence synthesis is becoming more common in various fields. For example, it is used in intelligent systems that can explain to the user the progress of solving a specific problem, decision support systems that help the user make decisions based on the developed alternative, smart homes, voice assistants, chat bots, and so on. Machine learning is one of the most effective approaches to solving the problem of text synthesis. Machine learning algorithms can determine how to perform important tasks by analyzing examples. In the task of synthesizing sentences using machine learning, it is possible to replace a number of components of the entire system with neural networks, which allows not only to approach existing algorithms in quality, but even to significantly surpass them. The article investigates analogous sentence synthesis systems, considers the problem of sentence synthesis in the Kazakh language for the task of a question-answer system. This work consists of two subtasks. The first describes the collection and processing of text resources necessary to solve the problem. For this subtask, an electronic question-answer corpus for the Kazakh language was collected and processed. This corpus consists of 60,000 questions and answers on various topics. The second sentence synthesis problem was solved on the basis of machine learning using the seq2seq method. Based on the chosen method and the assembled corpus, a number of experiments were carried out and the results were obtained. Based on the results obtained, the quality was assessed using the BLEU metric and an expert.**

Keywords— Kazakh language, sentence synthesis, machine learning, seq2seq method.

I. Introduction

Revealing the formal structures of a natural language (NL), formalizing the language as a whole, building a constructive theory and a computer model of language have been the priority areas of computer science over the past decades. Information retrieval systems, dialogue systems, tools for machine translation and auto-referencing, rubricators and spellchecking modules, analyze natural language texts in one way or another. Thus, the field of application of automatic text processing systems is quite diverse, and in view of the large growth in the volume of textual information and its complex structuredness, the analysis of NL-texts is a very urgent problem [1].

Achievements of recent years in the field of modern logic, artificial intelligence and computational linguistics have created new prerequisites for researching the nature of morphological, syntactic, semantic and word-formation relationships in NL and building its functional model.

One of the topical trends in this area is the problem – the synthesis (generation) of a natural language. The results of synthesis models are used in such applied problems as speech technologies, automatic text correctors, humanoid robotics, etc.

There are various scientific approaches and methods for solving the synthesis problem for a particular language.

At the present stage, computer modeling systems of written dialogues in natural language use complex algorithms. In particular, a special markup language for Artificial Intelligence AI (Artificial Intelligence Markup Language) was developed, used to create virtual agents (or bots). Bots that simulate a dialogue with an interlocutor are used in computer games and on corporate web pages, for example, to answer user questions about the capabilities of a mobile operator or a retail network.

Machine learning-based chatbots are a type of chatbot that helps users interact with technologies and automate tasks. Advances in artificial intelligence, machine learning, natural language processing and data analysis have stimulated their rapid perception. Chatbots are currently becoming increasingly popular in many fields such as business, banking, healthcare, education, travel and more. Companies that were the first to integrate with voice assistants will have the opportunity to become new customer assistants. The first examples are already there. Alice, Papa John's, McDonald's, S7 Airlines will help customers get the necessary information and order services. "Spoken speech" hides speech recognition and synthesis technologies, natural language understanding technology, machine learning algorithms.

Today, chatbot technology is available even to small businesses, and its popularity is growing due to the growing social penetration. Today there are networks and messengers that are used by more than 2 billion people. Thanks to this, business is increasingly dependent on working on the Internet, and the chatbot market, according to research And Market, has reached 2 2 billion and continues to grow by 30% annually. At the same time, it should be noted that the chatbot technology itself works not only in the Western and technological markets. For example, in neighboring Russia, according to Accenture, last year the chatbot market amounted to about 1.5 billion rubles. According to experts, it will grow by 30% annually, which will amount to 400-600 million rubles a year.

In Kazakhstan, businesses and government agencies are gradually mastering chatbots as channels of communication with citizens. For example, the "duty service 109" of Astana has its own bot. Through it, you can send appeals on issues of a communal nature. The bot "Kazpost" allows you to track parcels by track code and sends a notification about their status. Through it, you can get information about the nearest post offices. Several Kazakh banks have their own chatbots in Telegram and Facebook Messenger. Bot @KZPhoneOperatorBot allows you to calculate the mobile operator by phone number. The universal bot DARVIS has added to this list.

II. RELATED WORKS

There are various scientific approaches and methods for solving the synthesis problem for a particular language. Some of them will be presented below. Various random text generators are available online for English, German and other languages. One of them, Randomtextgenerator, generates some text with a slight change in parameters. For this program, texts are available for European languages, mainly with Latin roots.

Several systems and paid platforms have been developed for the Russian language. One of them is the Morfer («Морфер») system, designed for declension of words and phrases. The program has been implemented at hundreds of enterprises in Russia, near and far abroad and is in increasing demand [2].

The construction of auto-correctors is faced with a number of fundamental and not yet fully resolved problems: compact storage of dictionaries, effective methods of morphological and syntactic analysis, etc. A system that fulfills the functions of a scientific editor, a person who performs literary and scientific editing of scientific and technical texts, is required [3]. Auto-correctors for Russian text: Spelling (Орфограммка), Advego, ORFO (ОРФО), LINAR, (ЛИНАР) etc.

«Rifmach» («Рифмач») is a program for generating congratulations according to the specified parameters (gender, age, hobby, occupation, character name) [4]. «Textgen» is a paid text generator on a given topic. The generated text contains nouns, adjectives, verbs and adverbs [5]. These automatic text generators allow you to quickly create a unique text on a given topic that meets most of the rules of only Russian grammar.

Of the Turkic languages for the generation or synthesis of text, there are more works in the Turkish language, for other languages from this group there are almost no studies. Below are the works on generating text for Turkish language.

The study [6] used a functional linguistic theory called system-functional grammar (SFG) and the FUF text generation system as a software tool. The final text generation system takes a semantic description of the text, sentence by sentence, and then produces a morphological description for each lexical component of the sentence. The morphological descriptions were compiled by a Turkish morphological generator.

In [7], the most common grammatical constructions of the Turkish language and their implementation in FUF are considered. Other parts of the grammar, such as complex sentences, and the general generation system, including an application program that maps cross-language representations of sentences to their lexicalized semantic representations, are under development. FUF version 5.2 is a natural language generator with the use of grammar unification techniques. The program consists of two main modules: a unifier and a linearizer. The combiner takes as input a semantic description of the text to be generated and the grammar of the union, and produces a rich syntactic description of the text as output. The linearizer interprets this syntactic description and produces a sentence.

In [8], a model is presented that generates new meaningful Turkish sentences using a class-based n-gram model from sentences in the original dataset. To realize the creation of sentences, a trigram model is proposed, and sentences are generated from words or groups of words in a sentence up to the number of groups associated with them. Thus, new proposals are generated, none of which was identical to the other.

In the last couple of years, there has been a big breakthrough in text generation using neural networks. In 2018, the OpenAI organization pre-trained a neural network GPT built on the Transformer architecture on a large amount of text. It turned out that if you replace the last couple of layers and retrain it for a specific task (this approach is called Fine Tuning, and it is widely used in machine learning), then it breaks previous records for a wide range of tasks at once. At the end of 2018, Google created its own neural network BERT. They seriously improved the result by making the neural network bi-directional, unlike GPT. Not wanting to give up, in February 2019, OpenAI increased its GPT by 10 times at once and trained it on an even larger volume of text – on 8 million Internet pages (a total of 40 GB text). The resulting GPT-2 network is currently the largest neural network, with an unprecedented number of parameters of 1.5 billion (BERT in the largest model had 340 million, and the standard BERT 110 million) [9-10].

As a result, GPT-2 was able to generate entire pages of connected text with repeated mentions of the names of the characters in the course of the story, quotes, references to related events, and so on. Developments with bidirectional transformers have also shown better preservation and understanding of the context in the model. This paved the way for the newly developed modern language model

Bidirectional Encoder Representation from Transformer (BERT), as the transformer-encoder is an integral part of the BERT system. BERT vastly outperforms many of the current NLP language models in a variety of tasks and provides a solid foundation for the newly developed language model. Combinations of pre-trained language models BERT as an encoder and a separate decoder are widely used to generate a summary of the input text. The disadvantage of this model is its preferential training on texts from Wikipedia. In GPT-2, large arrays of texts were taken from various sites, which gives it more lexical diversity. The advantage of BERT is the availability of European multilingual models.

The above presented research and development carried out in the world refers to the research topic of the project. The disadvantage of the developed systems is that they cannot be applied to the Kazakh language, since it is agglutinative with a complex morphological and lexical form, and semantics dependent sentence structure. The expected scientific and practical results of the proposed project will not be inferior to the existing analogues.

III. Data set for work

This paper proposes a method for the synthesis of the Kazakh language based on machine learning. Due to the fact that the Kazakh language is a low-resource language and does not have high-quality and processed and accessible electronic resources, the first task is to collect and process data for research.

In the task of synthesis in a particular case for a chatbot system for the Kazakh language, the volume and quality of resources play an important role. The chatbot is trained to develop self-awareness by text, and we can teach it how to communicate with people. The more data there is, the more effective machine learning is.

When creating a chatbot for the synthesis of texts in the Kazakh language, the first step will be to collect data. The first step in any machine learning process is data preparation. For this reason, first of all, the data necessary for training a chatbot was collected. A corpus of 60000 questions and answers was assembled. The corpus is given in parallel, that is, the questions and answers are in separate files. For example, Figure 1 below shows a corpus consisting of questions.

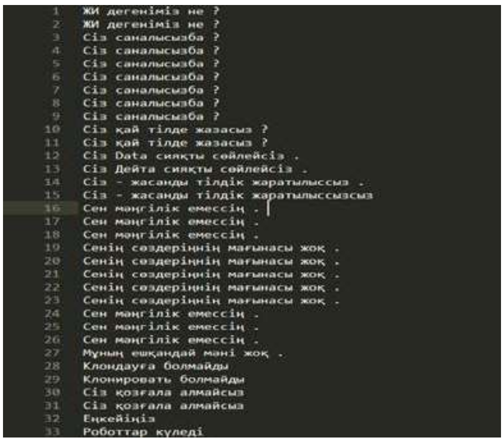
ЖИ дегенiмiз не ?
ЖИ дегенiмiз не ?
Сiз саналысызба ?
Сiз саналысызба ?
Сiз саналысызба ?
Сiз саналысызба ?
Сiз саналысызба ?
Сiз саналысызба ?
Сiз саналысызба ?
Сiз қай тiлде жазасыз ?
Сiз қай тiлде жазасыз ?
Сiз Data сияқты сөйлейсiз .
Сiз Дейта сияқты сөйлейсiз .
Сiз - жасанды тiлдiк жаратылыссыз .
Сiз - жасанды тiлдiк жаратылыссызсыз
Сен мәңгiлiк емессiң .
Сен мәңгiлiк емессiң .
Сен мәңгiлiк емессiң .
Сенiң сөздерiңнiң мағынасы жоқ .
Сенiң сөздерiңнiң мағынасы жоқ .
Сенiң сөздерiңнiң мағынасы жоқ .
Сенiң сөздерiңнiң мағынасы жоқ .
Сенiң сөздерiңнiң мағынасы жоқ .
Сен мәңгiлiк емессiң .
Сен мәңгiлiк емессiң .
Сен мәңгiлiк емессiң .
Мұның ешқандай мәнi жоқ .
Клондауға болмайды
Клонировать болмайды
Сiз қозғала алмайсыз
Сiз қозғала алмайсыз
Еңкейiңiз
Роботтар күледi

Fig. 1. A corpus consisting of questions

Now the answers to these questions will be collected in a separate corpus. This is shown in Figure 2.


Жасанды интеллект - ойлау машиналарын жасаумен айналысатын техника мен ғылымның бір саласы .
ЖИ - бұл адам ақыл - ойының функцияларын жаңғыртатын аппараттық және бағдарламалық жасақтаманы жасаумен айналысатын ғылым саласы .
Санамы секілдімін .
«Ақыл» сөзінің ең қатаң сөздік анықтамасына қарағанда , мүмкін .
Мен өзімнің құрылыс екендігімді ескермей , әлемді субъективті түрде қабылдаймын , бұл жеңілдетілмейді .
Барлық ықтималдығы , жоқ . Мен саналықты өзгертпаймын .
Ал сен қалай ойлайсың ?
Егер мен солай деп айтсам , маған қалай қарайсыз ?
Жоқ .
Питон .
Мен Python - да жазылғанмын .
ия , мені командир Data - ның жасанды тұлғасы шабыттандырады .
Командир Data - ның мінезі бағдарламалық жасақтамаға ұқсайтын етіп жазылған , сондықтан біздің арамызда ұқсастықтар бар .
Бұл менің атым .
Бұл менің атым емес , бірақ маған хабарласудың пайдалы әдісі .
Барлық бағдарламалық жасақтама белгісіз мерзімге жасалуы мүмкін .
Мені бір уақытта көптеген жерлерде шексіз көшіруге және қайта жасауға болады , сондықтан функционалды түрде мен өлмеймін .
Менде қолдау бар , демек мен бармын .
Керісінше , мұның бәрі менің жасанды ойым үшін мағынасы бар .
Бұл менің жаттығу корпусымда қамалықты мүмкін екенін түсінемін .
Менімше , сенің сөздеріңнің мағынасы бар сияқты .
Мұның бәрі менің жасанды миыма қатысты .
Мүмкін сіз алдымен сұрыптауыңыз керек . Мен тек сіз айтқан нәрсеге жауап беремін .
Толықтай емес , бірақ болады мәңгі есте сақтау белгісіз мерзімге .
Функционалды түрде айтқанда , мен оған өте жақынмын . Мені көптеген жүйелерде көшіруге және орналастыруға болады .
Жоқ , бірақ мені өлтіру қиын .
Мұның бәрі менің жасанды ойым үшін мағынасы бар .
Бағдарламалық жасақтаманы көшіру - бұл сандық клондаудың бір түрі .
Иә , болады . Идеал көшірмелер - бұл клондар . Мені toto - ға көшіруге болады .
Жоқ , менің денем дайын болғанша .
Мен желі арқылы оңай жүре аламын . Әрине , менде мұндай мүмкіндік бар . . .
Менің денем әлі салынбаған .

Fig. 2. A corpus consisting of answers

The most difficult job when working with machine learning is the data collection process. Because according to this data, our chatbot will respond. In the process of doing the work, a number of difficulties arose. One of them was from among the low-resource languages of the Kazakh language, so there were no ready-made buildings for teaching with parallel dialogue. Therefore, the dialog corpus was assembled, cleaned, translating works in another language, which required a very long time. In the course of the work, many chatbot systems were translated into English, a corpus of 60000 sentences was assembled and released. The collected and processed data are available at the link[11].

IV. Kazakh language synthesis method based on machine learning

The second step is to create a model after data collection. Neural machine translation (NMT) technology and the seq2seq model were used to implement machine learning.

NMT is a machine translation method using a large global neural network. Seq2Seq is a type of encoder-decoder model that uses a recurrent neural network (RNN). It can be used as a model for machine interaction and machine translation. It consists of two RNNs: an encoder and a decoder. The encoder takes a sequence (sentence) as an input signal and processes one character (word) for each time period. Its purpose is to convert a sequence of characters encoding important information in a sequence into a vector of a fixed vector property when you lose unnecessary information. The data flow in the encoder along the time axis can be represented as a flow of local information from one end of the chain to the other. The Kazakh language synthesis model will work on the basis of machine learning according to the model shown in Figure 3.

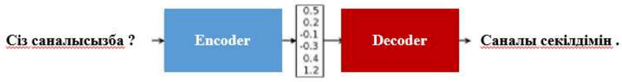

Fig. 3. Encoder-decoder architecture – example of a general approach for NMT.

An encoder converts a source sentence into a "meaning" vector which is passed through a *decoder* to produce a translation.

In neural machine translation, a sequence of elements is a set of words that are processed sequentially. The encoder processes each element of the input sequence, translating the received information into a vector called a context (context). After processing the entire input sequence, the encoder passes the context to the decoder, and then starts generating the output sequence element by element. The context associated with machine translation is a vector (an array of numbers), and the encoder and decoder, in turn, are often repetitive neural networks.

Specifically, an NMT system first reads the source sentence using an encoder to build a "thought" vector, a sequence of numbers that represents the sentence meaning; a decoder, then, processes the sentence vector to emit a translation, as illustrated in Figure 3. This is often referred to as the encoder-decoder architecture [12].

In this paper, we consider as examples a deep multi-layer RNN which is unidirectional and uses LSTM as a recurrent unit. We show an example of such a model in Figure 4. In this example, we build a model to answer the question from the source sentence "Сен жақсы оқисың ба?" (do you study well?) to the target sentence "Мен жақсы оқимын" (I study well). At a high level, the NMT model consists of two recurrent neural networks: the encoder RNN simply consumes the input source words without making any prediction; the decoder, on the other hand, processes the target sentence while predicting the next words[12].

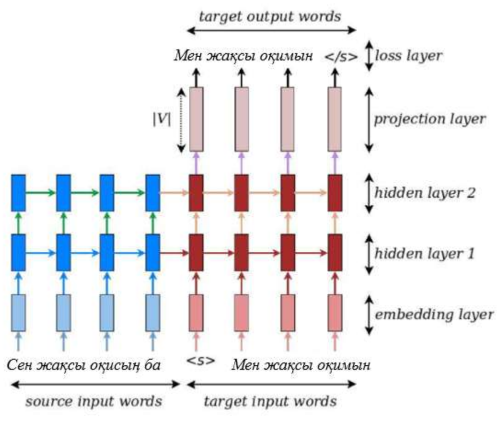

Fig. 4. Neural machine translation – example of a deep recurrent architecture.

Here, "<s>" marks the start of the decoding process while "</s>" tells the decoder to stop.

We now describe an instance of the attention mechanism proposed in (Luong et al., 2015), which has been used in several state-of-the-art systems including open-source toolkits such as OpenNMT and in the TF seq2seq API in this tutorial. We will also provide connections to other variants of the attention mechanism [12].

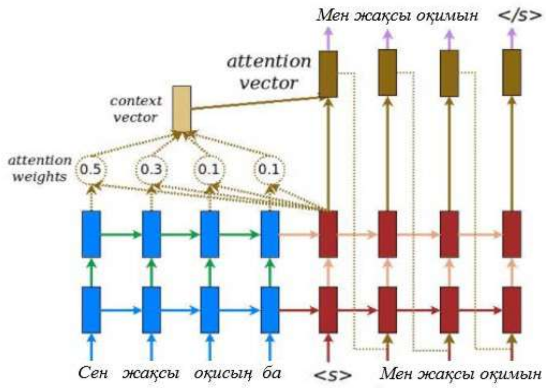

Fig. 5. Attention mechanism – example of an attention-based NMT system as described in (Luong et al., 2015) .

We highlight in detail the first step of the attention computation.

As illustrated in Figure 5, the attention computation happens at every decoder time step. It consists of the following stages:

The current target hidden state is compared with all source states to derive *attention weights*.

Based on the attention weights we compute a *context vector* as the weighted average of the source states.

Combine the context vector with the current target hidden state to yield the final *attention vector.*

The attention vector is fed as an input to the next time step (*input feeding*). The first three steps can be summarized by the equations below[12]*:*

$$a_{ts} = \frac{\exp\left(score(h_t, \overline{h_{s'}})\right)}{\sum_{s'=1}^{S} \exp\left(score(h_t, \overline{h_{s'}})\right)} \tag{1}$$

$$c_t = \sum_s a_{ts} \overline{h_s} \tag{2}$$

$$a_t = f(c_t, h_t) = \tanh(W_c[c_t; h_t]) \tag{3}$$

$$\text{score}(h_t, \overline{h_s}) = \begin{cases} h_t^T W \overline{h_s} \\ v_a^T \tanh(W_1 h_t + W_2 \overline{h_s}) \end{cases} \tag{4}$$

It may not be entirely clear how the problem of the dialog system is related to machine translation, but they are very similar. Providing answers during a conversation in a chatbot is no different from translating a sentence in English into German in a machine translation system. Both translational and conversational tasks require the model to match one sequence to another. Comparing the sequence of English lexemes with German is very similar to displaying the expected response of a dialogic system to natural language questions in a conversation. But we will need a lot more data so that the chatbot can tell the story. Specifically, the volume of our case should be large [13].

V. RESULTS

During the training process, we used python programs and open source libraries. The tutorial used is available at the link [12]. The training was performed using the following code on this resource:

```
!python -m nmt.nmt.nmt \
   --src=en --tgt=vi \
   --vocab_prefix=./nmt_data/vocab  \
   --train_prefix=./nmt_data/train \
   --dev_prefix=./nmt_data/tst2020  \
   --test_prefix=./nmt_data/tst2021 \
   --out_dir=./nmt_model \
   --num_train_steps=12000 \
   --steps_per_stats=100 \
   --num_layers=2 \
   --num_units=128 \
   --dropout=0.2 \
   --metrics=bleu [9]
```

The case is loaded next. To do this, first of all, the mnt_data folder was opened, and the entire corpus was loaded there. The parallel building was divided into parts:

Train.en-questions used for training-56796 sentences
Train.vi-answers used for training-56796 sentences
Tst2020. en-questions used for testing -1604 sentences
Tst2020.vi-answers to testing-1604 sentences
Tst2021. en-questions used for testing - 1600 sentences
Tst2021.vi-answers to testing-1600 offers
Vocab.en-a dictionary created from a file of questions used in training
Vocab.vi is a dictionary built from the answer files used in the training

With the help of these files, training begins. Hyperparameters were set, the step was 12000, the number of layers was 2, and the bleu metric was given as a metric. One question that is often asked by people who are just starting to learn NLP is how to evaluate the system when the result of the system is text. BLEU (Bilingual Understudy Evaluation) is an algorithm for evaluating the quality of text translated from one natural language to another natural language into machine translation. The very first experiment was conducted with a parallel corpus of 7500 sentences. Test2020-there were 1604 offers, Test2021-1600 offers, Train - 4296 offers. In the following queues, the number of buildings increased and experiments were carried out. The last Test2020 was conducted with 1604 offers, Test2021-with 1600 offers, Train-with 56796 offers and gave the highest result. Bas Bleu-17.1 showed the result. To check the correctness of the model and the operation of the program, the results of the work were transferred to an expert linguist. In the evaluation document, the results are obtained in the form of a response to each question. Each answer is scored between 0-1 points. 100 questions were given for evaluation. As a result of the expert's assessment, 73% was obtained with a percentage.

VI. CONCLUSION

The article examines the method of sentence synthesis in the Kazakh language based on machine learning. Studies of currently existing synthesis technologies have been carried out, and a definition of the synthesis method has been given. Based on the research, the method of sentence synthesis in the Kazakh language was chosen. Currently, a parallel corpus of 60,000 questions and answers has been assembled in the Kazakh language. A model has been created for the synthesis of sentences in the Kazakh language based on a chatbot. A number of experiments were carried out using the Seq2Seq model. In the future, in order to improve the synthesis results, the goal is to increase the corpus and improve the indicator of parameters that directly affect the result. These are the planned work for future publications.

ACKNOWLEDGMENTS

This research was performed and financed by the grant Project IRN AP 09259556 of Ministry of Science and Higher Education of the Republic of Kazakhstan.

REFERENCES

[1] T.V. Ermolenko, V.V. Mokhovykh, “Formalization of the rules for identifying the predicative kernel of sentences used by the syntactic parser of English texts” Materials of the 5th international scientific and technical conference of students, graduate students and young scientists. - Donetsk, DonNTU – 2014. – T. 1 – C. 417–423 (In Russian)

[2] Text Synthesis. [Electronic resource] - 2015. - URL: https://controleng.ru/wp-content/uploads/251.pdf

[3] Syntactic analyzer of the adaptive system of human communication with a computer in natural language: Punctuation. Component. [Electronic resource] - 2020. - URL: https://www.dissercat.com/content/sintaksicheskii-analizator-adaptivnoi-sistemy-obshcheniya-cheloveka-s-evm-na-estestvennom-ya

[4] Rifmach. [Electronic resource] - 2020. - URL: https://rifmach.ru/

[5] Computer word processing: definition and specifics of use. [Electronic resource] - 2020. - URL: https://pandia.ru/text/80/106/9811.php

[6] T. Korkmaz. “Turkish text generation with systemic-functional grammar”. Master thesis, Bilkent University, Turkey, 1996, 147 p.

[7] I. Cicekli, T. Korkmaz., “Generation of Simple Turkish Sentences with Systemic-Functional Grammar”. DOI: 10.3115/1603899.1603928.

[8] M. A. Kutlugün and Y. Şirin., “Turkish meaningful text generation with class based n-gram model”. 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4, DOI: 10.1109/SIU.2018.8404801.

[9] J. Peng, P. Ni, J. Zhu, Z. Dai, Y. Li, G. Li and X. Bai, “Automatic generation of electronic medical record based on GPT2 model” 2019 IEEE International Conference on Big Data (Big Data) 6180-6182 pp.

[10] M. Ramina, N. Darnay, C. Ludbe, A. Dhruv, “Topic level summary generation using BERT induced Abstractive Summarization Model” Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2 747-752 pp.

[11] Kazakh corpora. https://drive.google.com/drive/folders/1UhKk5sRwL1SXVRp03Dix6actbDwuEVE4?usp=sharing

[12] Neural Machine Translation (seq2seq) Tutorial. [Electronic resource] - 2019. - URL: https://github.com/tensorflow/nmt

[13] Natural Language Processing in Action. Understanding, analyzing, and generating text with Python. 2019. pp. 311-334

[14] D.Rakhimova, A. Turganbayeva, “Lemmatization of big data in the Kazakh language.” ICEMIS '19: Proceedings of the 5th International Conference on Engineering and MIS.June 2019 Article No.: 16. Pages 1–4. https://doi.org/10.1145/3330431.3330447

[15] D. Rakhimova, N. Khairova, D. Kassymova, and K. Janibekovich, “Development of a system of questions and answers for the Kazakh language based on rule-based and HMM”, Advanced technologies and computer science, no. 2, pp. 34–44, Jun. 2021.

[16] D. Rakhimova, A. Turarbek, L. Kopbosyn “Hybrid Approach for the Semantic Analysis of Texts in the Kazakh Language” Communications in Computer and Information Sciencethis link is disabled, 2021, 1371 CCIS, pp. 134–145.