The Task of Generating Text Based on a Semantic Approach for a Low-Resource Kazakh Language

Diana Rakhimova[1,2](✉), Satibaldiev Abilay[2], and Adilbek Kuralay[2]

[1] Institute of Information and Computational Technologies, St. Shevchenko 28, Almaty, Kazakhstan
di.diva@mail.ru

[2] Al-Farabi Kazakh National University, Al-Farabi 71, Almaty, Kazakhstan

Abstract. In this article, the authors consider the problem of text generation for low- resource languages, using the Kazakh language as an example, based on semantic analysis. Machine learning method is used in the generation of text documents and sources in the Kazakh language. First, semantic analysisis performed, the number of words in the given text, the number of stop words, the number of symbols, etc. Then the TF-IDF algorithm is used to find the semantically important words of the text. Annotation of the given text by means of semantic analysis. And at the end, generation of text with advanced semantic analysis. A corpus for the Kazakh language was prepared for experiments and research. GPT-3 and NLG are used in the process of generation. Generation by means of semantic analysis of the text gives us some great opportunities. The Recurrent Neural Network (RNN) method is used during generation.Generation gives us a lot of opportunities, including not spending time on unnecessary information. It will provide an article or short text related to the keywords you searched for. The description of the developed approach and practical results of experiments are presented.

Keywords: Semantic analysis · Machine learning · Text generation · RNN · Kazakh language

1 Introduction

The generation system for the Kazakh language developed on the basis of machine learning is one of the current issues. Using the generation system, we quickly and easily solve problems such as chat-bots, auto-abstract, writing poems, mathematical or geometric problems. Natural language processing problems are developing rapidly for the Kazakh language, and it should be said that the generation system in the Kazakh language is almost non-existent in the country at the moment. There are very few organizations working on these natural language processing reports, and those that exist do not publish their data Open Source. Therefore, it is our goal to solve this problem and publish it openly. Because if it is open, it will continue to develop and increase in data. And it will be affordable for small businesses as well, because it is financially inconvenient

N. T. Nguyen et al. (Eds.): ACIIDS 2023, CCIS 1863, pp. 585–595, 2023.
https://doi.org/10.1007/978-3-031-42430-4_48

for small businesses to develop technologies like chatbots. For private enterprises, this project will be of great benefit. Every web optimizer knows that a site must have unique texts in order to be liked by search engines. Not just any set of words, but meaningful sentences on the topic of the site. This is especially a problem for aggregators who receive information from other sites and online stores, where the parameters and data of the goods are usually the same. Therefore, standard practice in this case is to order unique texts from copywriters. Consider the task of automatically generating product descriptions based on reviews. Having multiple product user reviews from different sites, we automatically generate a small unique text that summarizes the information from the reviews. The large flow of information on the Internet has led to the rapid development of the natural language processing industry (NLP). Currently, various research mechanisms are developing their own projects, such as information exchange between users, machine translation of information, spam filters, e-mail verification and processing of question-and-answer systems. However, due to the lack of knowledge of the structure of some languages, there are problems where the research result does not fully meet the needs of the user. Today, one of the problems of search engines is the morphological and morphemic analysis of words encountered while processing user requests. An example of such languages is the Uzbek language, which belongs to the family of Turkic languages. Kazakh is one of the agglutinative languages. That is, in this language, each grammatical meaning is expressed by individual affixes. The term affix in the grammar of the Kazakh language is taken in the same general sense as in the grammar of other Turkic languages. This means prefixes, infixes, suffixes, conjunctions. Nowadays, the structure of the Kazakh language has become more complicated due to the influence of Arabic, Persian and Russian languages. Preprocessing input text data is a key initial step in any natural language processing (NLP) application. Extracting the base of the word, that is, extracting the base or root of the input word, is an important process in the preprocessing stage. That is, depending on the keywords you entered, a short answer or text will be generated. If the hulls have a large structure, the result will be a high structure. It is important to do the generation with high accuracy.

2 Related Works

Natural language processing is a powerful tool for creating a clear vision for the organization [1]. Application analysis of customer experience and activity social network helps the economic growth of the company [2]. However, sentimental analysis can lead to inaccuracies in reviews that include both positive and negative reviews [3]. This document focuses on the fact that the solution to this problem in the Kazakh language is still widely studied [4]. Recently, many researches have been conducted in the field of sentimental analysis in Indian, Arabic, Turkish languages [4–6], however, the number of researches is small for the Kazakh language [4, 13].

A study published in [5] used machine learning techniques for semantic analyses. Natural language support vector by training models with contract matching datasets machine (SVM), Naive Bayes. In addition, linguistic methods, such as the systematic use of special morphological analysis, have compiled sensory dictionaries of words and phrases, as well as a set of linguistic rules [4]. In addition, including pre-processing, morphological analysis techniques such as tokenization, word stopping, stamping and POS tagging in research [7] provide detailed information about the data for high accuracy in the results. Evaluation of reviews using semantic analysis created a pattern of neural bags resulting from negative or positive reviews. Word bag model is a method of performing textual data in the process of text modeling with machine learning algorithms. Bag-of-words model is not complex and advances in problems such as device and seen language modeling and document classification [14]. If we consider the problem of generation after semantics.

Text generation is one of the popular problems in data science and machine learning and is suitable for Recurrent Neural Networks. This report uses TensorFlow to create an RNN text generator and create a high-level API in Python3. The solution to the problem was inspired by the work of Aurelien Heron [8]. This CST463 is a great project at Cal State Monterey Bay's Advanced Machine Learning Program led by Dr. Glenn Bruns [9].

Recurrent Neural Network (RNN). A real limitation of vanilla neural networks (as well as convolutional networks) is that their APIs are limited: they take a fixed-size vector as input (like an image) and produce a fixed-size vector as output (like probabilities of different classes). And not only that: these models perform this comparison using a fixed number of computational steps (such as the number of layers in the model). The main reason recurrent networks are interesting is that they allow us to work with sequences of vectors: sequences in the input, sequences in the output, or, in general, both [10].

Natural language generation (NLG) is a subfield of natural language processing. NLG focuses on some basic semantic representation of information from written text generation in natural languages. NLG is used in many applications: Multilingual reporting, text summarization, machine translation, and dialog applications. Therefore, the automated production of language is associated with a large number of diverse theoretical and practical problems. In NLG systems, problems such as multi-content selection, text-based lexicalization, text integration, and linking expressions are common.

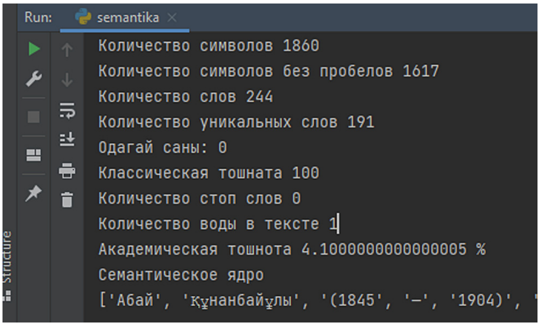
Natural language text generation is a recommended way to introduce communication. Semantic graphs are the most representative systems used as input to NLG [9]. Among them, due to the limitation of the representation of the semantic graph, the traditional type of operational or procedural knowledge is incomplete, so it is necessary to assign more structure to the nodes, as well as links [11]. In natural language text generation, there was a great need for more rich graph detail. A new semantic representation called Rich is a Semantic Graph (RSG) that contains additional information. The main purpose of this stage is to evaluate and then arrange the paragraphs according to two factors: consistency between paragraph sentences and synonyms of the most frequently used paragraph words. After experimental testing, we found that the coherence measure produces very close results, so synonyms of the most frequently used paragraph words are used as an additional evaluation factor. First, text consistency assessment is used to assess whether paragraphs are consistent or not.

Therefore, each paragraph is evaluated and ranked according to the number of coherence between its sentences. Second, the synonyms of the most frequently used paragraph words are collected by entering the WordNet rank. Finally, the last paragraphs can be sorted according to the relevance rating, followed by the most frequently used paragraph word degree of synonyms [5]. After that, it will be more efficient to generate with semantics. Adding semantics will help us get better results as the generation produces words or texts that are semantically relevant to the keywords you are looking for. Extracts matching text from corpora based on semantics.

At the moment, many Turkic languages, like the Kazakh language, are of low resource. Due to the lack of available linguistic resources, it is difficult to apply modern methods and develop high-quality technologies in the field of NLP and artificial intelligence.

3 Description of the Model

We conducted research related to our topic and created a semantic analysis model for the Kazakh language through generation. The semantic model consists of three parts. The first is to produce statistics of the text in the Kazakh language. We have researched all sections and prepared the practical section. We will share the results of the semantic generation model for the Kazakh language below. We took out the statistics of the text presented in the first part, entered a short text. As a result, we calculated the number of words, symbols, punctuation marks, punctuation marks in the text (Figs. 1 and 2).



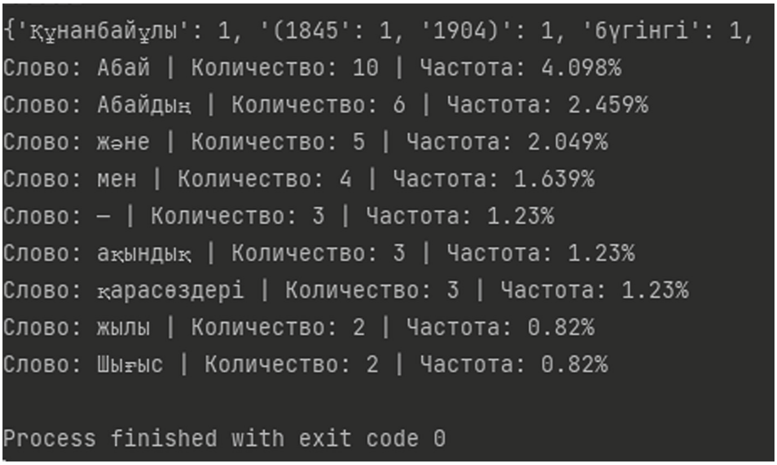
```
Количество символов 1860
Количество символов без пробелов 1617
Количество слов 244
Количество уникальных слов 191
Одагай саны: 0
Классическая тошната 100
Количество стоп слов 0
Количество воды в тексте 1
Академическая тошнота 4.1000000000000005 %
Семантическое ядро
['Абай', 'Құнанбайұлы', '(1845', '–', '1904)', '
```

Fig. 1. Example of statistical data of the semantic analysis of the text in the Kazakh language.

Figure 1 shows the total number of words, the number of unique words, classic nausea, academic nausea, semantic core-keywords of the text.

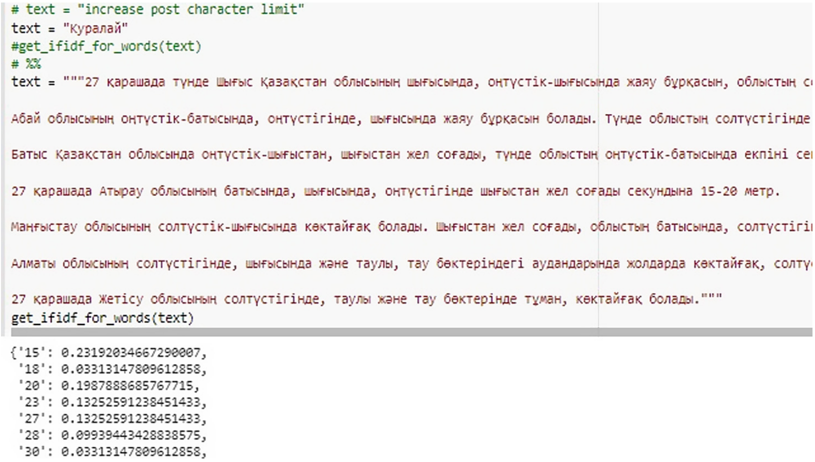
In the second part, we used TF-IDF to extract semantically important sentences for a known Kazakh language text. In this part, we used TF-IDF in order to further improve the semantic analysis [15]. Given a certain text, determine the frequency of each sentence of that given text. Finds the importance of sentences for the text and extracts the most important sentences with their weights (Fig. 3 and 4).

We have generated this pre-text below (Fig. 5 and 6). After testing our model, we gave it the word "science" as an input, and we got the result, which can be seen in (Fig. 5). As a result, we studied only one scientific text, which is not of high quality. But this problem will be solved in the future, because we have a database of more than 35 million



```
{'Құнанбайұлы': 1, '(1845': 1, '1904)': 1, 'бүгінгі': 1,
Слово: Абай | Количество: 10 | Частота: 4.098%
Слово: Абайдың | Количество: 6 | Частота: 2.459%
Слово: және | Количество: 5 | Частота: 2.049%
Слово: мен | Количество: 4 | Частота: 1.639%
Слово: – | Количество: 3 | Частота: 1.23%
Слово: ақындық | Количество: 3 | Частота: 1.23%
Слово: қарасөздері | Количество: 3 | Частота: 1.23%
Слово: жылы | Количество: 2 | Частота: 0.82%
Слово: Шығыс | Количество: 2 | Частота: 0.82%

Process finished with exit code 0
```

Fig. 2. Number and frequency of each word relative to the text.



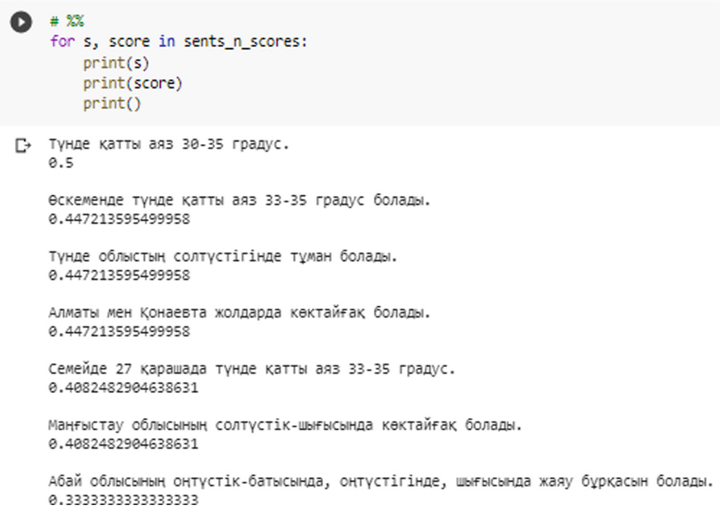
```
# text = "increase post character limit"
text = "Куралай"
#get_ifidf_for_words(text)
# %%
text = """27 қарашада түнде Шығыс Қазақстан облысының шығысында, оңтүстік-шығысында жаяу бұрқасын, облыстың с

Абай облысының оңтүстік-батысында, оңтүстігінде, шығысында жаяу бұрқасын болады. Түнде облыстың солтүстігінде

Батыс Қазақстан облысында оңтүстік-шығыстан, шығыстан жел соғады, түнде облыстың оңтүстік-батысында екпіні се

27 қарашада Атырау облысының батысында, шығысында, оңтүстігінде шығыстан жел соғады секундына 15-20 метр.

Маңғыстау облысының солтүстік-шығысында көктайғақ болады. Шығыстан жел соғады, облыстың батысында, солтүстігі

Алматы облысының солтүстігінде, шығысында және таулы, тау бөктеріндегі аудандарында жолдарда көктайғақ, солтү

27 қарашада Жетісу облысының солтүстігінде, таулы және тау бөктерінде тұман, көктайғақ болады."""
get_ifidf_for_words(text)
```

```
{'15': 0.23192034667290007,
 '18': 0.03313147809612858,
 '20': 0.1987888685767715,
 '23': 0.13252591238451433,
 '27': 0.13252591238451433,
 '28': 0.09939443428838575,
 '30': 0.03313147809612858,
```

Fig. 3. Meaning of words for text according to TF-IDF of semantic analysis.

words collected from Kazakh-language web pages. But first, before using this data, we need to prepare our model for retraining, otherwise we run the risk of memory overflow if we feed all the data to our model [12]. Therefore, we use the Gradient Optimization method to solve this problem (Fig. 5).

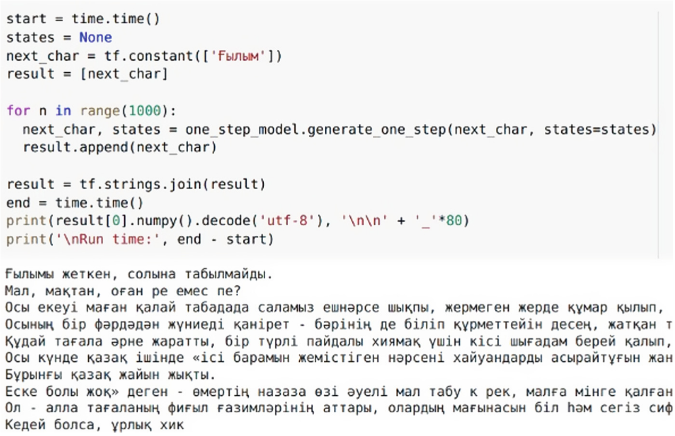
In order to test how well our model is learning, we develop the prediction truth and error metrics, which can be seen in Fig. 6 below. As we can see in the figure, the truth prediction metric has a maximum value of 0.5, which means that the sequence of symbols does not have a high probability of placement. To increase the accuracy, we need to retrain the model and train the model by adding new data. The x-axis here is from 0–200, and it is known that our model consists of 4,070,247 parameters, dividing



```
# %%
for s, score in sents_n_scores:
    print(s)
    print(score)
    print()
```

```
Түнде қатты аяз 30-35 градус.
0.5

Өскеменде түнде қатты аяз 33-35 градус болады.
0.447213595499958

Түнде облыстың солтүстігінде тұман болады.
0.447213595499958

Алматы мен Қонаевта жолдарда көктайғақ болады.
0.447213595499958

Семейде 27 карашада түнде қатты аяз 33-35 градус.
0.4082482904638631

Маңғыстау облысының солтүстік-шығысында көктайғақ болады.
0.4082482904638631

Абай облысының оңтүстік-батысында, оңтүстігінде, шығысында жаяу бұрқасын болады.
0.3333333333333333
```

Fig. 4. Meaningfulness of sentences for text according to TF-IDF of semantic analysis.



```
start = time.time()
states = None
next_char = tf.constant(['Ғылым'])
result = [next_char]

for n in range(1000):
  next_char, states = one_step_model.generate_one_step(next_char, states=states)
  result.append(next_char)

result = tf.strings.join(result)
end = time.time()
print(result[0].numpy().decode('utf-8'), '\n\n' + '_'*80)
print('\nRun time:', end - start)
```

```
Ғылымы жеткен, солына табылмайды.
Мал, мақтан, оған ре емес пе?
Осы екеуі маған қалай табадада саламыз ешнәрсе шықпы, жермеген жерде құмар қылып,
Осының бір фәрдәдән жүниеді қанірет - бәрінің де біліп құрметтейін десең, жатқан т
Құдай тағала әрне жаратты, бір түрлі пайдалы хиямақ үшін кісі шығадам берей қалып,
Осы күнде қазақ ішінде «ісі барамын жемістіген нәрсені хайуандарды асырайтұғын жан
Бұрынғы қазақ жайын жықты.
Еске болы жоқ» деген - өмертің назаза өзі әуелі мал табу к рек, малға мінге қалған
Ол - алла тағаланың фиғыл ғазимләрінің аттары, олардың мағынасын біл һәм сегіз сиф
Кедей болса, ұрлық хик
```

Fig. 5. The result of text input to the model for generation.

those parameters by 200 and reading it, the prediction probabilities of the symbol at each stage (Fig. 6).

The result of the adjustment model for the generation model of the Kazakh language is as follows (Table 1). Since the model is trained using the recurrent training method, the result of the model does not give significant semantic value, but it has the correct structure of the word. This model has an error rate of 20%, which means that even if the model cannot form a clear idea, it will write words correctly and try not to deviate from the topic.

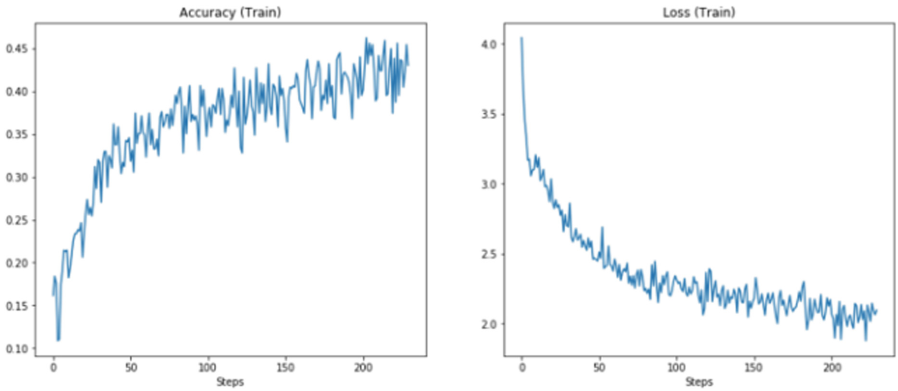


Fig. 6. Metrics indicator of the model.

Table 1. The result of the adjustment model for the generation model of the Kazakh language.

Data type	Example of experimental results
Input text (in Kazakh)	*Қазақстан республикасы*
Output text (in Kazakh)	Қазақстан Республикасы, Шығыс Еуразияда орналасқан елдің өзіндік республикалық жүйесімен басқарылатын мемлекет. Қазақстан Республикасының ең кіші мемлекеттік бірлігін құрастырушы мекеме Парламент, Президент және Ұлттық сәбиекорлық суды. Қазақстан, Қазақстан жерінің шекарасы, Тянь-Шань тасмағының ортақтамасы және Орталық Азияда орналасқаны арқылы бұрынғы шамамен Гүней Еуразияда орналасқан биік алқабылардың бірі. Қазақстан елінің оңтүстігінде, Сырдария және Есіктің кеңістігінде, Каспий теңізінен, Жетісудан және Алтын Емелден бастап, Қарағанды түсіндігіне дейінгі аумақты қамтылады Қазақстан Республикасы, жер орынды дайындық, топографиялық, климаттық, экологиялық, өнеркәсіптік-санақтардың орналасуы және жер-жер аймақтарының жаттығуына көмек көрсету әрекеттерін жүргізеді
Input text (English translation)	*Republic of Kazakhstan*

(*continued*)

Table 1. (*continued*)

Data type	Example of experimental results
Output text (English translation)	The Republic of Kazakhstan is a country in Eastern Eurasia governed by its own republican system. The constituent institutions of the smallest state unit of the Republic of Kazakhstan are the Parliament, the President and the National Child Protection Court. Kazakhstan is one of the high valleys located in South Eurasia, due to the border of the land of Kazakhstan, the border of the Tien-Shan belt and its location in Central Asia. In the south of the country of Kazakhstan, in the space of the Syrdarya and Esik, the territory from the Caspian Sea, Zhetysu and Altyn Emel to Karaganda basin is included The Republic of Kazakhstan conducts activities to assist in the preparation of land, topographical, climatic, environmental, industrial-census location and training of land-regions
Input text (in Kazakh)	*Коронавирус салдары*
Output text (in Kazakh)	Коронавирус салдары дегеніміз, Сарсылмақтан көмекшеленген ауруларды аталады. Коронавирус (COVID-19), Сарсылмақтан қозғалған өзгеріп жататын коронавирустың жаңа түріне байланысты ауру. Осы ауру 2019 жылы Китайда пайда болған және оның көлігінде жалғасты. COVID-19 оқиғасы ғана Китайда шығып, кейін сонымен байланысты басқа өлкелерге жетті Коронавирус аурусы тамақ қауіпсіздігін, көтерілген жастарын, диабетпен, асқазан сигаретті қатталдықпен, бүйір ауруларымен, бет-бетше оралып жатқан жағдайда байланыстырады. Онда көптеген кезде жас, көп адамдар ауру тым өте жасасып, дәрігерлерге медициналық көмек көрсетуі қажет болады Коронавирус салдарының көбіне қанаттар арқылы туындауы, көзқарастар арқылы ауысуы және сауаттылығын қорғау үшін қалаулар қолданылады. Жаттатын гигиена әдістері мен әуелгі ортақ маска носу арқылы жасампаздықты қамтамасыз ету мүмкін. Бірнеше өлкелердегі жергілікті ауружайлары қорына көтерілген COVID-19 вакцинасылары да қолданылады
Input text (English translation)	*Consequences of the coronavirus*
Output text (English translation)	The consequences of the coronavirus are called measles-assisted diseases. Coronavirus (COVID-19), a disease caused by a new type of mutated coronavirus transmitted by Sarsylmak. This disease appeared in China in 2019 and continued in his car. The case of COVID-19 only started in China and then spread to other countries The coronavirus disease is associated with food insecurity, elevated youth, diabetes, stomach stiffness, side effects, and face-to-face contact. In many cases, many young people will become very sick and need medical help from doctors The consequences of the coronavirus are often generated through wings, shifting through attitudes and preferences are used to protect literacy. Creativity can be ensured by practicing hygiene techniques and wearing the first shared mask. In several regions, local hospitals stockpiled COVID-19 vaccines are also used

The table shows fragments of the result of generating text in the Kazakh language. You can see not a bad result. The text is relevant to the topic and has grammatically structured sentences.

4 Evaluation and Discussion

To evaluate the quality of the algorithm for each class, we calculate the Precision and Recall metrics separately. Precision can be interpreted as the proportion of objects called positive by the classifier that are also truly positive, and recall indicates what proportion of objects of the positive class the algorithm found among all objects of the positive class. There are several different ways to combine precision and recall into a summary quality measure. Table 2 presents the results of evaluating the quality of the text generation model on given corpora.

Table 2. The evaluation of the obtained results of the text generation model testing in the Kazakh language

Name of the corpus	Source	Number of characters	Recall	Precision
Health	https://kitaphana.kz	2067887	61.65	69.97
Republic of Kazakhstan	https://bankreferatov.kz	2360983	72.74	75.18
Historical figures	https://bankreferatov.kz	6154140	69.35	73.85

For the experiment, the names of the requests were dependent on the topics and genres of the corpora. Additionally, the requests consisted of 1 to 3 words. The average results obtained from the testing were as follows: Recall = 67.9, Precision = 73. The quality of the results is not satisfactory. During the experiment, the lowest results were observed in text generation for the literary and scientific genres. This was due to the specific themes and structural forms of the texts themselves. For the scientific genre, only scientific articles were considered, which limited the model's learning process. To resolve and improve the quality of the model, future plans involve increasing the quantity and quality of the corpora. However, using all available data for the model without proper cleaning beforehand may lead to overfitting and compromise its performance. To address this issue, future plans involve the utilization of the Gradient Descent method to optimize the cleaning process and enhance the model's overall performance.

5 Conclusion

The developed model of Kazakh language text generation based on semantic analysis presents not bad results. However, it is difficult to achieve such accuracy, which is not 100% accurate. But the more the trained information structure, the higher the result can be achieved. Semantic analysis of the Kazakh language compared to other languages is somewhat difficult. The lack of information and data and the complexity of the morphology of the Kazakh language have a somewhat negative effect. Digital data in the Kazakh language have been collected and supplemented. A prototype of the generation system model for the Kazakh language was created and we trained the model on the collected corpus. The created model was tested and discussed. Recurrent Neural Networks (RNNs) have been studied and discussed. Information about the linguistic resources of the Kazakh language was analyzed. We focused these studies mainly on semantics and generation. Our main task was to generate semantic text. We have developed this model. In the course of this research, corpora in the Kazakh language were collected and models were created. According to the results of the research, search for a text, article or text from the corpus related to the keywords you entered or searched for, and produce the text that is close to the semantics. That is, firstly, it saves time, and secondly, getting rid of unnecessary information. Since generation is important, it is used in various spheres. For example: chatbot, search engines, Q&A in many companies. It allows to save the budget, reduce time, and reduce the number of workers.

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. Yemm, G.: Can NLP help or harm your business? (2006)
2. Ranjan, S., Sood, S., Verma, V.: Twitter sentiment analysis of real-time customer experience feedback for predicting growth of Indian telecom companies. In: 2018 4th International Conference on Computing Sciences (ICCS). IEEE (2018)
3. Liu, B.: Sentiment analysis: a multi-faceted problem. IEEE Intell. Syst. **25**(3), 76–80 (2017)
4. Yergesh, B., Bekmanova, G., Sharipbay, A.: Sentiment analysis on the hotel reviews in the Kazakh language. In: 2017 International Conference on Computer Science and Engineering (UBMK). IEEE (2017)
5. Phani, S., Lahiri, S., Biswas, A.: Sentiment analysis of tweets in three Indian languages. In: Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2016) (2016)
6. Baly, R., El-Khoury, G., Moukalled, R., Aoun, R., Hajj, H., Shaban, K.B., El-Hajj, W.: Comparative evaluation of sentiment analysis methods across Arabic dialects. Procedia Comput. Sci. **117**, 266–273 (2017)
7. Yildirim, E., Çetin, F.S., Eryigit, G., Temel, T.: The impact of NLP on Turkish sentiment analysis (2016)
8. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. CST 463-Advanced Machine Learning. https://catalog.csumb.edu/preview_course_nopop.php?catoid=1&coid=476. Accessed 10 Nov 2022
9. Sherstinsky, A.: Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) Network. https://arxiv.org/abs/1808.03314. Accessed 28 Oct 2022
10. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: Kremer, S.C., Kolen, J.F. (eds.) A Field Guide to Dynamical Recurrent Neural Networks. IEEE Press (2001)
11. Haber, E., Ruthotto, L.: Stable architectures for deep neural networks. Inverse Probl. **34**(1), 014004 (2017)
12. Rakhimova, D., Turarbek, A., Kopbosyn, L.: Hybrid approach for the semantic analysis of texts in the Kazakh language. In: Hong, T.-P., Wojtkiewicz, K., Chawuthai, R., Sitek, P. (eds.) ACIIDS 2021. CCIS, vol. 1371, pp. 134–145. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-1685-3_12
13. Diana, R., Assem, S.: Problems of semantics of words of the Kazakh language in the information retrieval. In: Nguyen, N.T., Chbeir, R., Exposito, E., Aniorté, P., Trawiński, B. (eds.) ICCCI 2019. LNCS (LNAI), vol. 11684, pp. 70–81. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28374-2_7

14. Rakhimova, D., Turganbayeva, A.: Approach to extract keywords and keyphrases of text resources and documents in the Kazakh language. In: Nguyen, N.T., Hoang, B.H., Huynh, C.P., Hwang, D., Trawiński, B., Vossen, G. (eds.) ICCCI 2020. LNCS (LNAI), vol. 12496, pp. 719–729. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63007-2_56
15. Rakhimova, D., Turganbayeva, A.: Auto-abstracting of texts in the Kazakh language. In: Proceedings of the 6th International Conference on Engineering & MIS, pp. 1–5 (2020)