IRSTI 16.21.25

A. Bilakhanova[1], A. Ydyrys[2], N. Sultanova[3]
[1,2,3]Suleyman Demirel University, Kaskelen, Kazakhstan

KAZAKH LANGUAGE-BASED QUESTION ANSWERING SYSTEM USING DEEP LEARNING APPROACH

Abstract. Deep learning advances have resulted in considerable gains in a variety of natural language processing applications, including question-answering (QA) systems. QA systems are intended to retrieve data from big datasets and respond to user queries using natural language. Deep learning-based techniques have yielded encouraging results in the development of QA systems capable of providing consistent answers to a wide range of inquiries. This research presents a deep learning-based Kazakh language-based QA system. A pre-processing module is also included in the proposed system to improve the quality of the input text and the accuracy of the final output. The results reveal that the system has a high level of accuracy. This study promotes to the advancement of question-answering technology and contributes to the development of natural language processing tools in the Kazakh language.

Keywords: Kazakh language, question-answering system, natural language processing, deep learning approach, accuracy.

***

Аңдатпа. Терең оқытудағы жетістіктер табиғи тілді өңдеуге арналған әртүрлі қосымшаларда, соның ішінде сұрақтарға жауап беру жүйелерінде айтарлықтай жетістіктерге әкелді. Сұрақтарға жауап беру жүйелері үлкен деректер жиынтығынан деректерді алуға және табиғи тілді қолдана отырып, пайдаланушылардың сұрауларына жауап беруге арналған. Терең оқытуға негізделген әдістер сұрақтардың кең ауқымына дәйекті жауап беруге қабілетті сұрақтарға жауап беру жүйелерін әзірлеуде жігерлендіретін нәтижелер берді. Бұл зерттеуде терең оқытуға негізделген қазақ тіліндегі сұрақтарға жауап беру жүйесі ұсынылған. Алдын ала өңдеу модулі кіріс мәтінінің сапасын және соңғы шығыс дәлдігін жақсарту үшін ұсынылған жүйеге енгізілген. Нәтижелер жүйенің жоғары дәлдік деңгейіне ие екенін көрсетеді. Бұл зерттеу сұрақтарға жауап беру технологиясын дамытуға ықпал етеді және қазақ тілінде табиғи тілді өңдеу құралдарын әзірлеуге өз үлесін қосады.

Түйін сөздер: Қазақ тілі, сұрақ-жауап жүйесі, табиғи тілді өңдеу, терең оқыту тәсілі, дәлдік.

***

Аннотация. Достижения в области глубокого обучения привели к значительным достижениям в различных приложениях для обработки естественного языка, включая системы ответов на вопросы. Системы ответов на вопросы предназначены для извлечения данных из больших наборов данных и ответа на запросы пользователей с использованием естественного языка. Методы, основанные на глубоком обучении, дали обнадеживающие результаты в разработке систем ответов на вопросы, способных давать последовательные ответы на широкий спектр запросов. В этом исследовании представлена система ответов на вопросы на казахском языке, основанная на глубоком обучении. Модуль предварительной обработки также включен в предлагаемую систему для улучшения качества входного текста и точности конечного вывода. Результаты показывают, что система обладает высоким уровнем точности. Это исследование способствует развитию технологии ответов на вопросы и вносит свой вклад в разработку средств обработки естественного языка на казахском языке.

Ключевые слова: Казахский язык, вопросно-ответная система, обработка естественного языка, подход к глубокому обучению, точность.

I. Introduction

The question-answering (QA) system is an intelligent program activated by natural language input. They provide dialogue output in response. Although the system was originally designed for entertainment purposes, with the advancement of data mining, machine learning, and deep learning technologies, the system is becoming more popular and useful in education, business, and among other industries. Deep learning methods have demonstrated their effectiveness in numerous natural language processing (NLP) tasks, and this has contributed to the emergence of question-answering systems that rely on these techniques to generate precise and dependable answers to challenging queries. Such systems are able to leverage extensive amounts of structured and unstructured data in order to learn and improve their performance.

For agglutinative languages, it is crucial to implement such a type system where words are formed by adding morphemes, which can change the meaning of the word. Although there have been efforts to develop multilingual models, there's a lack of data and expertise in training these algorithms for different languages. The Kazakh language's agglutinative nature, complex derivational structure, and rich morphology present challenges for NLP. Thus, the purpose of this study is to explore the obstacles in the construction and focus on the implementation of a model of a QA system, based on the Kazakh language using deep learning methods.

II. Literature review

Question-answering systems have been a subject of research for decades, and recent advancements in deep-learning neural networks have significantly improved the performance of these systems. The use of deep learning models has enabled question-answering systems to process large volumes of data and generate accurate answers to questions in natural language. This literature review explores recent studies in the development of QA systems based on deep learning techniques.

One popular approach in developing question answering systems is to use neural networks, particularly deep learning neural networks. The authors of a research paper [1] discussed various Deep Learning algorithms offered in the field of Question Answering and evaluated their performance on twenty tasks from Facebook's babI dataset. The report also contained implementation details and algorithm enhancements developed to generate better outcomes.

The most prominent deep learning-based approach for QA is the Transformer architecture [2]. The model is entirely based on attention processes, with multi-headed self-attention replacing the conventional recurrent layers found in encoder-decoder designs. The Transformer has served as a core model for numerous cutting-edge natural language processing models, including question answering systems.

The researchers presented a multi-stage neural network model for answering questions in a follow-up paper [3]. Known as Bi-Directional Attention Flow (BIDAF), the hierarchical neural network can represent the interactions between a query and a passage. The BIDAF model underwent testing on the SQuAD dataset and exhibited state-of-the-art performance.

Another approach in developing question answering systems is to use pre-trained language models. In a paper [4], the authors proposed a language model called BERT (Bidirectional Encoder Representations from Transformers) that is capable of pre-training on large volumes of text and fine-tuning on downstream tasks such as question answering. The BERT model achieved state-of-the-art performance on the SQuAD dataset and has since been used in various question answering systems.

Agglutinative Languages:

Agglutinative languages present unique challenges for NLP, as each morpheme can change the word's meaning. Therefore, it is crucial to develop NLP systems that can handle these languages. Researchers have attempted to develop language-specific models for such languages. For instance, Turkish researchers [5] developed QA system for Turkish that employs sophisticated AI algorithms and large datasets, particularly the BERT algorithm, to generate language models and a fine-tuning mechanism for machine reading in QA tasks. The system selects the best answer from a collection of banking-related documents. The research team conducted several experiments using both original and translated data sets to address the difficulties presented by Turkish's complex morphology. Mongolian researchers [6] presented a new approach for the Mongolian question-answering system known as Interactive Mongolian Question Answer Matching Model (IMQAMM), which utilizes an attention mechanism. The IMQAMM is composed of two key components: interactive information enhancement and max-mean pooling matching. In interactive information enhancement, sequence enhancement and multi-cast attention are utilized to create scalar features via numerous attention mechanisms. Mongolian morpheme representation is also integrated into the model to aid with semantic feature learning. The model was evaluated on a Mongolian corpus of question-answer pairs related to the law domain. The experimental outcomes demonstrate that the IMQAMM model outperforms the baseline models significantly.

In the case of the Kazakh language, which is an agglutinative language with a complex derivational structure and rich morphology, there is limited research on QA systems. Researchers have attempted to develop Kazakh language models for other NLP tasks, such as sentiment analysis and text classification. For instance, in one study [7] proposed a Kazakh text classification system using convolutional neural networks (CNNs), while another study [8] created a character-based language model for the Kazakh language using Deep Neural Networks, specifically the Long Short-Term Memory model, to generate correct words given context. To assure accuracy, the model was trained using Kazakh-language texts. In another study [9] proposed a multi-task learning model called MTQU (Multi-Task Query Understanding) that utilizes deep learning techniques and features unique to the Kazakh language to establish connections between tasks such as question categorization and named entity recognition. The model also has a multi-feature input layer, which is shown to have a positive effect on training performance. Results from experiments demonstrate that the MTQU model effectively enhances question classification and named entity recognition.

While deep learning neural networks have shown great promise in developing question-answering systems, there are still challenges that need to be addressed. One challenge is the lack of annotated datasets for training and evaluating these systems. Another challenge is the interpretability of these models, as it is often difficult to understand how the model arrived at a particular answer.

Agglutinative languages present unique challenges for NLP, and researchers have attempted to develop language-specific models for such languages. There are still challenges that need to be addressed and there is a lack of research on QA systems in Kazakh, but the promise of these models in answering questions in natural language is a promising direction for future research. Therefore, this paper focuses on implementing QA system model for the Kazakh language using deep learning techniques.

III. Method

A. Data Description

To train the model the babI dataset (released by Facebook) has been used. The dataset (11K) was translated into Kazakh and saved in a CSV file. The columns in the dataset include:

- <Story (S)>: a statement regarding the facts relevant to the situation in question.
- <Question (Q)>: a question related to the story.
- <Answer (A)>: an answer to the question in the form of yes/no.

B. Preprocessing

In this part, we process Stories and Questions to extract keywords. The processing of the text begins by removing punctuation and then applying a morphological analyzer [11] to stem the text, resulting in the removal of suffixes from words. The extracted keywords were utilized in further stages of the study.

The initial challenge in utilizing analytical tools on text data is to convert it into a form that computer systems can understand and manipulate, as they only process numbers. A common solution to this problem is to represent words as vectors, as they have a meaningful interpretation and can be applied to various tasks [12].To convert stories, questions, and answers into numerical sequences with equal lengths, we determined the maximum length for stories and questions. We iterated over each story, question, and answer in the data, converting the raw words into numerical values and adding each set to its respective output list. Then, we padded the sequences to ensure they were all the same length. The result of this process is a tuple of the vectorized stories, questions, and answers (S, Q, A).

C. System Architecture

In 2015 Sukhbaatar et al. [10] proposed a neural network model called the End-to-End Memory Network (MemN2N), which is designed for QA tasks that require reasoning over long-term memory. The MemN2N model uses a combination of neural network layers, including an input module, a memory module, and an output module, to store and retrieve information from a memory matrix. The model is trained end-to-end to learn how to perform question answering tasks based on the provided input and output examples. The MemN2N model has demonstrated effectiveness on several benchmark datasets and has been shown to outperform existing approaches to problems requiring complex reasoning.

Embedding Layer: An embedding layer is used to convert the input vectors into continuous representations. This allows the model to learn more complex relationships between the words.

Input Module: The input module takes the embedded input vectors and generates a set of memory vectors that represent the story. This is done by iteratively updating the memory vectors based on the input vectors.

Question Module: The question module takes the embedded question vector and generates a query vector that is used to search the memory vectors for the answer. This is done by calculating a similarity score between the query vector and each memory vector.

Output Module: The output module takes the similarity scores and generates a probability distribution over the possible answers. This is done using a softmax function.

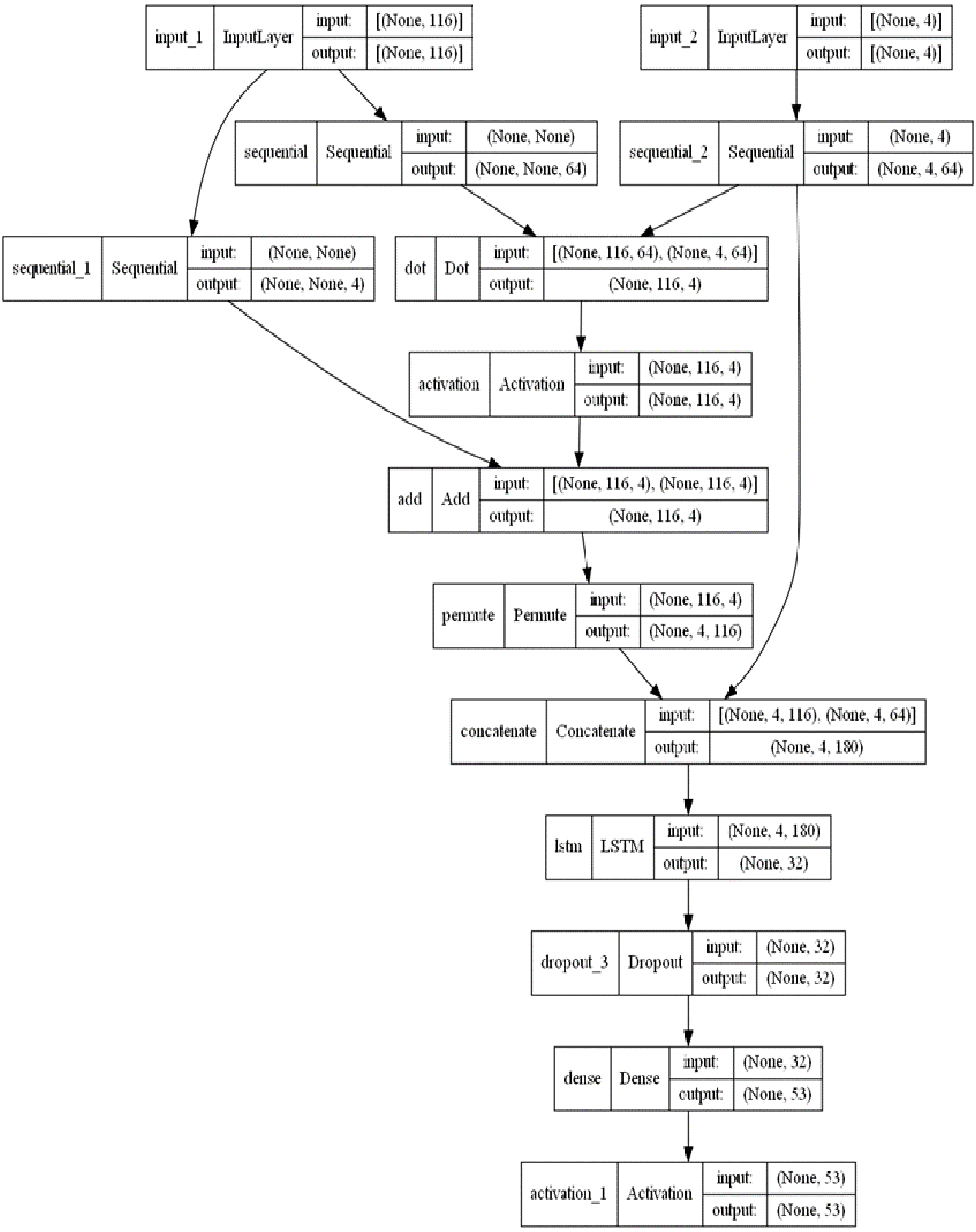
input_1 | InputLayer | input: | [(None, 116)]
output: | [(None, 116)]
input_2 | InputLayer | input: | [(None, 4)]
output: | [(None, 4)]
sequential | Sequential | input: | (None, None)
output: | (None, None, 64)
sequential_2 | Sequential | input: | (None, 4)
output: | (None, 4, 64)
sequential_1 | Sequential | input: | (None, None)
output: | (None, None, 4)
dot | Dot | input: | [(None, 116, 64), (None, 4, 64)]
output: | (None, 116, 4)
activation | Activation | input: | (None, 116, 4)
output: | (None, 116, 4)
add | Add | input: | [(None, 116, 4), (None, 116, 4)]
output: | (None, 116, 4)
permute | Permute | input: | (None, 116, 4)
output: | (None, 4, 116)
concatenate | Concatenate | input: | [(None, 4, 116), (None, 4, 64)]
output: | (None, 4, 180)
lstm | LSTM | input: | (None, 4, 180)
output: | (None, 32)
dropout_3 | Dropout | input: | (None, 32)
output: | (None, 32)
dense | Dense | input: | (None, 32)
output: | (None, 53)
activation_1 | Activation | input: | (None, 53)
output: | (None, 53)

Figure 1. End-to-end Memory Neural Network

The model takes a discrete set of inputs (real sentences or stories) {X_1, ..., X_n,} which are going to be stored in the memory, a question Q, and an answer A. Each of the Xi, Q, and A carries symbols from a lexicon with a large number of V words. A V is one that contains the entire vocabulary across all datasets. After writing all X into memory up to a preset buffer size, the model finds a continuous representation for the X and Q. Following that, output A is processed using the continuous representation after a number of hops. Due to multiple memory accesses, this enables backpropagation of the error signal to the input during training (Figure 1).

IV. Results

We have trained the model on different batch sizes, epoch, and dropout. The batch size is from 36 to 100 and the epoch is 20-150, the dropout is 0.2-0.5. The model showed constant accuracy, not less than 80%. We got higher accuracy with batch size = 36, epoch size = 150 and dropout = 0.2. The higher accuracy of the model was 92%, F1 score 93% as well. In the Figure 2 below, you can see that the model has reached 80% accuracy at epoch = 20.

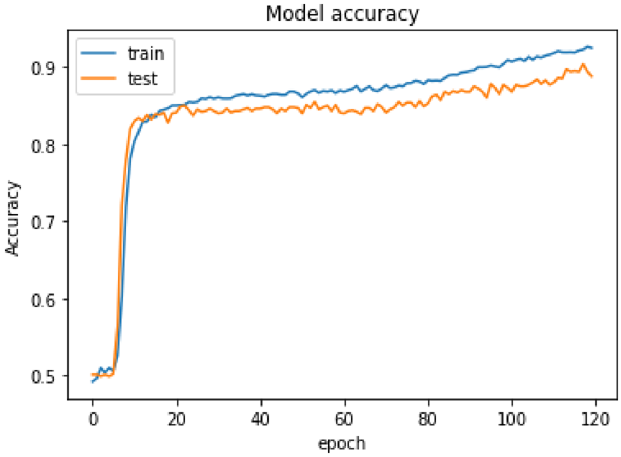

Figure 2. Model accuracy

After achieving improved accuracy, we tested the model and found that it was able to answer questions from the test data with a confidence of 99%. To validate its performance, we also tested the model with a custom story and question that included words present in the vocabulary used in the dataset. The results are displayed in Table 1.

Table 1. Results of Test data and Custom data

	Test data	Custom data
Story	"Маржан сүтті сонда алды. Жахан жатын бөлмеге бет алды. Маржан сүтті тастады. Жахан бақшаға кетті."	"Жахан ас үйге бет алды. Аяжан футбол добын бақшаға тастады."
Question	"Жахан ас үйде ме?"	"Футбол добы бақшада ма?"

True Answer	"жоқ"	"иә"
Predicted Answer	"жоқ"	"иә"
Probability of certainty	99%	88%

The analysis of the findings shows that the system properly answers the majority of the questions where the response is a single fact. Overall, the model answered 922 of 1000 questions correctly. The confusion matrix is given below:

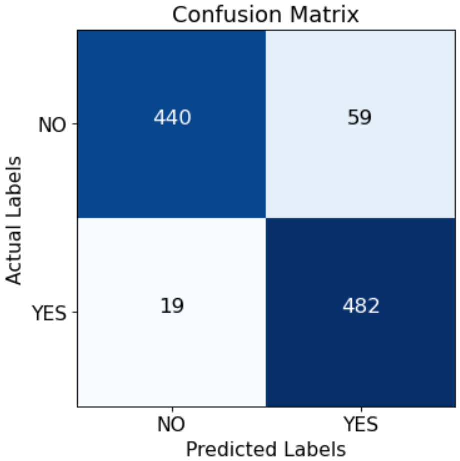

Figure 3. Confusion matrix of the Model

V. Conclusion

In conclusion, the objective of this paper was to implement a deep learning model for Kazakh language in the domain of Question Answering systems. To accomplish this, a translated version of the babI dataset from Facebook was utilized to train the model. The results were impressive, with the model demonstrating an accuracy rate of 92% and an F1 score of 93%. These outcomes are a clear indication of the efficacy of the methods used, along with the comprehensive preprocessing carried out on the dataset, which factored in the intricacies of the Kazakh language.

While the model has demonstrated promising results, there is still significant potential for improvement. In our future endeavors, we intend to augment and refine our dataset further. Additionally, we plan to create a user-friendly and accessible GUI application for our system to improve its usability.

References

1 Sharma Y., Gupta S. Deep learning approaches for question answering system. Procedia computer science. 2018 Jan 1;132:785-94.

2 Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.

3 Seo M., Kembhavi A., Farhadi A., Hajishirzi H. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603. 2016 Nov 5.

4 Devlin J., Chang M. W., Lee K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.

5 Gemirter C.B., Goularas D. A. Turkish question answering system based on deep learning neural networks. Journal of Intelligent Systems: Theory and Applications. 2021 Sep;4(2):65-75.

6 Yutao P., Weihua W., Feilong B. Interactive Mongolian Question Answer Matching Model Based on Attention Mechanism in the Law Domain. In Proceedings of the 21st Chinese National Conference on Computational Linguistics. 2022; (pp. 896-907).

7 Parhat S., Ting G., Ablimit M., Hamdulla A. A morpheme sequence and convolutional neural network based Kazakh text classification. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2019 Nov 18; (pp. 1903-1906). IEEE.

8 Sultanova N., Kessikbayeva G., Amangeldi Y. Kazakh Language Open Vocabulary Language Model with Deep Neural Networks. In 2019 15th International Conference on Electronics, Computer and Computation (ICECCO). 2019 Dec 10; (pp. 1-4).

9 Haisa G., Altenbek G. Multi-Task Learning Model for Kazakh Query Understanding. Sensors. 2022 Dec 14;22(24):9810.

10 Sukhbaatar S., Weston J., Fergus R. End-to-end memory networks. Advances in neural information processing systems. 2015;28.

11 O. Makhambetov, A. Makazhanov, I. Sabyrgaliyev, Zh. Yessenbayev. "Data-driven morphological analysis and disambiguation for Kazakh". In International Conference on Intelligent Text Processing and Computational Linguistics 2015, pp. 151-163.

12 Almeida F., Xexéo G. Word embeddings: A survey. arXiv preprint arXiv:1901.09069. 2019 Jan 25.