ABSTRACT

of dissertation work by Sultanova Nazerke Zholdybaevna, "Open Vocabulary Model for Kazakh Language using Deep Neural Networks", submitted for a PhD degree in specialty 6D070400 - "Computing systems and software"

General description of the work. This model is devoted to developing a generative model for Kazakh language with the application of neural networks. In the work two models are proposed: pure character-based model and character-based model enhanced with the attention layer. Deep Neural Networks were applied for constructing generative language model for Kazakh. The language modeling is recognized as sequence prediction task. Keeping the long-range dependencies is essential since the aim of the work is to generate valid words according to the context given. Taking that into account, recurrent neural network named long-short term memory was adopted. In order to enhance the performance of text understanding, the attention layer has been added for the For text generation purposes and detection of new words, the model needs to understand the word structure and syllable and character sequences. To be successful in this task, the book “Abay Zholy” by Mukhtar Auezov that was originally written in Kazakh was used for training purposes.

Assessment of the current state of the scientific and technological problem being solved. For the past 25 years there has been a demand for software solutions related to text processing, which has repeatedly experienced periods of growth, related to the emergence of personal computers, and with the rapid development of the Internet, and the rapid development of the Internet, and, In this natural language remains the most important way of communication, be the input of the search query on the miniature screen of the mobile phone, hints of the car navigator or business correspondence. Practically in all such applications such or otherwise the language model is used. So, for a convenient input of texts on a mobile phone, it is necessary to use the predicate input system, which practically corresponds to the direct application of the language model; language model - an indefinite part of the system of speech recognition, including volume and vocal search; Linguistic models are used in machine translation systems, the quality of which at the moment is still far from ideal, but still grows steadily.

Relevance of the research topic. Natural language processing helps computers communicate with people in their native language and scale other language tasks. For example, NLP allows computers to read text, hear speech, interpret it, measure mood, and determine which parts are important. Modern machines can analyze more language data than humans, without fatigue and in a consistent, unbiased manner. Given the vast amount of unstructured data that is generated every day, from medical records to social media, automation will be critical to efficiently analyzing text and speech data.

While digitalization takes place in all places of services in Kazakhstan, natural language processing is an indispensable part of this process. Taking into account that the natural language understanding field is still in developing state for Kazakh language,

building a language model seems to be a good start to realize the digital projects in Kazakh language.

Natural Language models tend to be the vital tools in computational linguistics. The task of language modeling is to determine the probable distribution of words over the chains of words in some language. The language modelling task appears in such practical areas as speech recognition, optical recognition of symbols (OCR), recognition of handwritten text, machine translation, spelling check, predicate input and the rest.

The purpose of the study is the development of a language model for Kazakh text generation using deep neural networks technologies. The objective of the work is to develop the character-based generative language model for the Kazakh Language. Language model in this research is a sequence to sequence generation task, where an input is a set of the words, and output is the batch of the words which are constructed using characters. A word can be remedied as a morpheme produced by characters in which any attainable word option may be produced.

The target is to deliver unseen words which might be easily fit into the particular context. Here comes the addition of attention layer for the LSTM model to enhance the syntactic performance of the work.

Research tasks, their place in the implementation of research work in general. To achieve the planned results of work, the following tasks have been identified:

- study and analyze the current state of the art of language models for different languages
- to develop a functional diagram and architecture of recurrent neural model
- to develop methods and algorithms for character-based language modeling using recurrent neural networks
- analyze and justify the choice of optimization models for the text generation models
- compare the performance of the developed model with the state of the art

Research methods. The main research approaches are methods of computational intelligence. Deep Neural Networks were applied for constructing generative language model for Kazakh. The language modeling is recognized sequence prediction task. Keeping the long-range dependencies is essential since the aim of the work is to generate valid words according to the context given. Taking that into account, recurrent neural network named long-short term memory was adopted. LSTMs are specifically designed to address long-term dependencies problems. Their specialization is the storage of information for long periods of time, so they practically do not need to be trained. All recurrent neural networks are in the form of a chain of repeating modules of a neural network. For these tasks, LSTM models were used to generate words because of their ability to remember the previous state.

Based on an abstract model, a neural network was developed. The vector model of the language is used as a coder, which allows moving towards greater "meaningfulness" of the model and "understanding" of the meaning of words. A recurrent neural network

acts as a decoder, since it allows processing information cyclically as it moves from input to output, and the output depends on previous calculations, providing a "memory" effect. The network architecture has three layers: the first layer is the attention mechanism, and two layers in each the input word layer, the projection layer, the recurrent layer, and the softmax layer. Adding an attention mechanism helps the decoder make better use of the input text information.

The scientific novelty of the research topic is determined by the fact that the innovative language model has been built. The research towards Open Vocabulary Language Model for Kazakh language has been conducted in this work. The use of neural networks is essential to overcome the sparseness problem and produce relevant results. Moreover, the character-based neural model is suggested to compromise with limitedness of vocabulary. Therefore, the mentioned goal will be achieved by implementing by stage and analyzing the proposed models into a language. Moreover, results shed light on future tasks to emerge the type of words in a context by adding word type information also. This can be very useful in agglutinative languages as the Kazakh language where the meaning and structure of words are very dependent on word endings.

The substantive novelty differs from the previous models based on the application of the neural networks using the graphical processing unit that makes the computation more efficient. The constructed architecture with attention layer enhanced the model performance.

The developed model can preserve the logic of the narrative and build dialogues on relatively short texts (3-4 sentences long), but it lacks a global context and preservation of the structure of the narrative, as in real works of art.

Provisions for Defense. The following provisions are submitted to the defense:

- methods and algorithms for language modeling and text generation;
- methods and algorithms for Kazakh language model;
- novel network architecture to generate Kazakh text;
- results of experiments and discussion

The structure and scope of the thesis.

The dissertation work consists of normative references, list of symbols and abbreviations, an introduction, 5 chapters, a conclusion, a list of references. It is presented on 82 pages of typewritten text, contains 21 figures, 11 tables, a list of used sources of 108 titles.

The work starts with an introduction where the author gives a general description of the work, novelties and provisions for the defense. In chapter 1, the previous work towards language modelling has been discussed. Huge amount of literature review has been conducted towards text generation for different languages. Chapter 2 covers a review of language models, namely the statistical approach and the approach based on artificial neural networks. Chapter 3 describes experiments with character-based model, based on a long short-term memory network. The results for the character-based model and first system architecture is presented in this chapter. In chapter 4, natural language

processing tasks are shown where language modelling can be applied: including part of speech analysis, word stemming, text classification and sentiment analysis. In chapter 5, an innovative model is described, using a deep neural network with an attention mechanism, for Kazakh language. General overview of the attention model has been presented and an application for language modelling with the proposed output has been written.

Within the framework of this dissertation work, 11 research papers on the topic under consideration were prepared and published, of which:

- three article each in a foreign publication and in an international peer-reviewed scientific journals;
- four articles were published in publishing houses that meet the requirements of the highest certification commission of the Ministry of Education and Science of the Republic of Kazakhstan;
- four articles have been published in the proceedings of international conferences.

The main scientific results of the dissertation research, set out in the dissertation, are presented at the international scientific conference in Nigeria, and the results are published in IEEE Xplore proceedings:

Kazakh Language Open Vocabulary Language Model with Deep Neural Networks. 2019 15th International Conference on Electronics, Computer and Computation (ICECCO), Abuja, Nigeria, 2019, pp. 1-4, doi:10.1109/ICECCO48375.2019.9043253.