ANNOTATION
of the thesis of Akanova Akerke Saparovna
"Neurocomputer system of semantic analysis of the text",
presented for the degree of Doctor of Philosophy (PhD)
in the specialty: 6D075100 Informatics, Computer Science and Control

Relevance of the work

Kazakhstan is one of the dynamically developing countries, which strives to match the level of the global educational space. The coronavirus pandemic, which forced all representatives of the educational society to turn to information and communication technologies, also made its adjustments to the global educational space. The pandemic made educational society communicate through computer technology and had thereby shown the value of the rapid exchange of information. The most common type of information is information presented in the form of texts in the language of certain country. In Kazakhstan it's Kazakh language. Kazakh language belongs to the Turkic languages of the Kipchak group, a type of synthetic agglutinating languages. The words in it consist of a stem and affixes added to it. There may be several affixeswhich allow tomake up to 60 new words from one word.

Hence, automatic processing of the text in the Kazakh language for the purpose of semantic analysis will allow to select the most essential, conceptual points important for this particular task from the set of texts and will reduce the teacher's work effort for checking the text works. Nowadays the Kazakh language in computer linguistics has achieved great results. Studies of scientists in the field of automation of text processing in the Kazakh language using a neural network deserveattention butconsidered insufficient. There is still lack of models of text processing in the Kazakh language without morphological analysis, without ontological dictionaries, which has been studied and covered in this paper.

Thus, the topicality of research posed by task of creating new models of classification by the content topic of Kazakh texts, the solution of which allows you to achieve a qualitative determination of the semantic proximity of the text content to a given topic.

The purpose of doctoral thesis

The purpose of the thesis is to create a model and algorithms used to solve the problem of determining the semantic proximity of the text content to a given topic by deep learning of thematic model output, which contributes to the creation of a neurocomputer system for semantic analysis of texts in the Kazakh language.

In order to achieve this purpose, the following *tasks* are carried out as part of the study:

1) Research and analyse semantic text analysis methods and models used in automatic text processing

2) Analyse technology that facilitates deep learning of neural networks.

3) Develop a dictionary of word forms for automatic processing of Kazakh

texts

4) Develop an affix truncation algorithm for Kazakh words

5) Develop a thematic model of documents in Kazakh language

6) Develop a multilayer neural network for training the result of a thematic model

7) Develop a neurocomputer system for semantic analysis of the text.

Research methods

In the course of the study various methods were used to solve the set tasks. Methods of automatic text processing for semantic analysis, methods of analytical research and mathematical statistics, including methods of statistical hypothesis testing (statistical criteria) based on the Student distribution were widely used. The inverse error propagation method, as well as computer modelling tools were used to build a mathematical model of deep neural network training. In the process of the experimental part of the work automation tools of mathematical calculations and tools for visualization of results based on Python programming language were used.

The object of research in this thesis is automatic processing of Kazakh texts**.**

The subject of the research in this thesis is the Classification of the content of Kazakh texts by topic.

The scientific novelty of this study is to develop a neural network model in computer linguistics research, requiring to determine the accuracy of the semantics of the text used, based on the algorithm of word affix truncation to create a dictionary of word forms when solving the problem of classifying Kazakh texts.

Authenticity and justification. The authenticity and justification of the results are confirmed by a comprehensive analysis of the literature on the topic of research and the conducted neural network training experiment. All factors were taken into account when calculating the probability of the distributions of topics in the document and words in the topics, the error function in the neural network training, the learning optimizer, and the metric.

The authenticity of the performed work is justified by the publication of the research findings in a peer-reviewed journal and the implementation of the neurocomputer system in the educational process.

The theoretical significance is in the application of the stemming algorithm and neural network model in research related to Kazakh text processing.

The practical significance is in the application:

– an affix truncation algorithm for Kazakh words, which is used to create a wordform dictionary for thematic modelling (LDA);

– neural network model with 4 layers for deep learning of LDA-model output data:

а) optimal algorithms (layers) of the neural network were selected;

б) optimal parameters for neural network compilation, which affect the result of neural network training, were determined.

– a neurocomputer system for semantic analysis of Kazakh language text material based on the thematic model and the built neural network model.

Algorithm of truncation of affixes of Kazakh words and the trained neural network got its practical implementation in the neurocomputer system, which is used for semantic analysis of text works in Kazakh language.

Provisions for the defence:

– algorithm for truncating affixes of Kazakh words;

– thematic model of documents in Kazakh language.

– neural network model for training semantic analysis vectors of Kazakh texts;

–deep neural network training

– a neurocomputer system for analysing Kazakh texts and probabilistic determination of the semantic proximity to the given topic.

Approbation of the work. Approbation of the work took place at the base of S. Toraighyrov Pavlodar State University and an act of implementation was obtained. The main publications were presented at international scientific conferences, published in scientific journals, the developed electronic products were registered in the state register of rights to objects protected by copyright:

- Materials of the Republican Scientific-Practical Conference "Innovations in Professional Education: Problems and Prospects" dated for celebration of 80th anniversary of Professor Bolat Abdikarimuly, February 2019, Astana.

- Technology Audit and Production Reserves. DOI: 10.15587/2706-5448.2020.217613

- Certificates of data inclusion in the state register of rights to objects protected by copyright № 8300 of February 20, 2020 and № 18050 of May 27, 2021.

The results of the dissertation were published in 10 papers. Among them: 1 monograph, 3 articles in the journals recommended by the Committee for Control of Education and Science of the RoK MES, 1 article in a domestic scientific journal, 1 article in an international scientific journal included in the database Scopus, 1 article in a foreign scientific journal, 1 paper in the proceedings of international and national conferences, 2 certificates of copyright.

Author's personal contribution.

The main experimental and theoretical results obtained in the course of the thesis research were obtained by the author independently. The degree-seeking student is the main author who owns the main ideas in obtaining, summarizing and analyzing the results achieved in the publications developed by team of co-authors. The structure of the thesis includes the following sections: introduction, main part (three chapters), conclusion, list of references and appendices. The work contains 113 pages of computer text, 30 figures, 15 tables and 199 titles of bibliographic sources.

Main results of the study.

As a result of the study of scientific sources is a **classification of methods** used in automatic text processing (Figure 4) in general.

It describes **method of error propagation** in deep neural teaching, which was applied in the development of a neurocomputer system of semantic analysis of

Kazakh.

A comparative analysis of object recognition technologies based on neural networks was conducted, which described comparative characteristics of 13 technologies– Apache Singa, Caffe, Deeplearning4j, Keras, Microsoft Cognitive Toolkit, MXNet, Neural Designer, OpenNN, Theano, Torch, Tensor Flow, NLTK, Gensim. Scientists are trying to map the structure of the human brain through artificial neural networks (ANNs). The above analysis provides that it is more convincing to use Gensim technology to classify text data. The other technologies are better suited for image, video and sound recognition.

An affix truncation algorithm was developed on basis ofPorter's stemmer to create a reference wordform dictionary. The creation of wordforms can be done by different methods, but the most common one is stemming. Stemming is the process of extracting the stem of a word by dropping endings and suffixes. Python programming language has implemented a stemmer to truncate affixes in Kazakh texts.

Thematic model is built using LDA to process the text in Kazakh language. Creattion of thematic model needs the use of the corpus of texts in Kazakh language, which consists of a collection of documents. The corpus of texts is then tokenized, which gives us a division of the text into words. With the help of a stemmer, suffixes and endings are seperated from the stem and a dictionary is created. Words from the resulting dictionary are extracted using bigrams, which determine the sign of the proximity of words to each other, after obtained results bigrams define topics using bag-of-words (BOW) technology. This work used ready-made algorithm doc2bow, which finds the keywords approximated by meaning. Then using the LDA algorithm keywords are allocated to topics and topics are allocated to documents. As a result of LDA we get keywords with weights relative to topics, distribution of topics with weights by documents.

The result of the obtained LDA model is evaluated by performing deep learning with the built neural network model. Thus, a neural network with 4 layers was built for training. The input layer Embedding() - converts keywords with weights and topics with weights into vector data. The second layer SpatialDropout1D () - makes regularization of neurons in the network. The third layer LSTM contains two more regulating layers inside: one regular dropout, and one recurrent dropout. The fourth layer is the output Dense layer. Each layer performs its function with respect to the received data in the neural network. Each layer is an algorithm for processing neurons by calculating a weighted sum. The learning process begins with back error propagation - the process of calculating the received error in the opposite direction, that is, from the output to the input.

A neurocomputer system for processing text data in the Kazakh language is developed. The stages of development of the neurocomputer system are given:

1. Creating a corpus of Kazakh texts.

2. Creating a dictionary of stems by applying the affix truncation algorithm for the Kazakh language.

3. Update the dictionary with selected bigrams or unigrams relative to topics (keywords).

4. Creating a topic model (LDA).
5. Creating a multilayer neural network.
6. Multilayer neural network training.
7. Linking the task and the trained neural network model.
8. Creating an interface for human-machine interaction.
9. Testing of neurocomputer system for semantic analysis of Kazakh texts.

The architecture of the neurocomputer system of semantic analysis of the text is developed. Testing of neurocomputer system is performed by the verification of compliance of the Kazakh text material content to the subject material. At the experimental level, correlation coefficient was calculated, which showed a linear dependence between the results of the experts and the program itself, which confirms the reliability of the developed models and the trained neural network. An act of software implementation at the Toraighyrov University was obtained.