ANNOTATION

of Bekarystankyzy Akbayan

on the specialty 8D06103 – "Management information systems" on the topic "Development of end-to-end system for automatic recognition of speech in agglutinative languages"

Relevance of the research topic. Automatic Speech Recognition (ASR) systems are nowadays widely used in different areas of human life, in order to make it easy for people to interact with computer systems and different applications. For example, smart assistants, smart home systems, commercial and subtitling applications allow to control computer systems without touching, from a distance. Moreover, ASR can make it easy to impaired persons to use electronic devices. For example, group of scientists study the ways of building ASR systems for people with dysarthria. Dysarthria is the type of muscle defects responsible for articulation. ASR development for people with this problem helps them interact not only with digital systems but also with other persons. Next useful example of involving ASR systems is the assessment of hearing loss. This type of system can forecast the level of hearing injury by the quality of answers to questions. But these opportunities are available only for people who knows widely used languages, like English, Chinese and Russian. ASR development for low-resource languages still needs enormous efforts, like data collection and preparation, testing of well-known recognition architectures, as well as studying the ways of adjusting state-of-the art architectures for exact languages or the group of languages. The Turkic group of agglutinative languages, to which the Kazakh language belong to, has many low-resource languages. Besides the problem of shortage of data to train, agglutinative languages have other problems stated out below.

Development of ASR systems for agglutinative languages are complex processes due to the their morphological complexity and richness of grammatical forms in these languages. According to this, development and fine-tuning ASR systems for agglutinative languages require additional studies and specific approaches. Below is a list of several challenges ASR systems for agglutinative languages can face:

- Morphems' analysis and separation. Morphemes in agglutinative languages can be joined and can be complex, which makes difficult the process of splitting and analysis of morphemes in speech recognition.
- Variety of word-formation rules: in agglutinative languages usually there exist various rules of word formation which determine how to

concatenate affixes with the root of a word. This also requires complex models and rules for processing those rules in ASR systems.

There is an enormous number of studies dedicated to the development of specific approaches and models to get reliable ASR systems for agglutinative languages. Some authors propose a language model, based on morphemes, where morphemes are understood as any of prefix, root or suffix in a word. As a result, authors got an automatic speech recognition system with large vocabulary. One research studies the performance of transformer-based CTC system which depends on context, trained with the word pieces taken as training units. Authors note the effectiveness of their method not only for English and German, but also for one of agglutinative languages - Turkish language. Next study made research on applying transformer architecture for a morphological disambiguator using Turkish language. This disambiguator can be used in any of NLP tasks and speech recognition is not exception here. The transformer architecture performed well also for another agglutinative language - Hindi. Here the transformer architecture along with Connectionist Temporal Classification (CTC), Language Model (LM) showed the lowest error rate for Hindi language: 3.2%. One more example of using Transformer architecture in ASR development for agglutinative language is for Finnish language. Here the author compares the performance of Transformer-XL architecture with LSTM and concluded that perplexity improvement achieved 29% and Word Error Rate (WER) was decreased to 3% for Finnish language. In the next paper authors state out that there are the most widely used and effective end-to-end architectures for automatic speech recognition: connectionist temporal classification and attention-based mechanism. Also, in this work mentioned the lack of transcribed audio-text pair resources for agglutinative languages to train in order to develop reliable automatic speech recognition systems.

According to the mentioned researches for agglutinative languages it was noted that dictionary enlarging and transformer architecture are the most effective approaches for developing end-to-end automatic speech recognition systems for agglutinative languages. Moreover, the lack of data to train and common morphological rules and similar soundings of languages from Turkic family of agglutinative languages served as a basis for providing pooling experiments, like transfer learning and multilingual training for these languages.

Purpose of the dissertation. The present dissertation was developed with the aim of studying the ways of improving ASR performance for agglutinative languages on the example languages from Turkic family.

Research objectives.

1) Analysis of existing ASR approaches for general cases and for agglutinative languages.
2) Extension and development of data corpus for agglutinative languages.

3) Development of models and methods for automatic speech recognition of agglutinative languages.
4) System development for automatic recognition of speech in agglutinative languages.

Object of the study. Modern automatic speech recognition methods and approaches, especially pooling methods like multilingual training and transfer learning.

Subject of the study. Agglutinative languages of Turkic family, methods of Machine learning, namely neural networks for Automatic Speech Recognition: attention mechanism, convolutional neural networks, performance improvement methods for critically low-resource languages, a moment from Natural Language Processing methods: word embeddings, and demonstrative Telebot which uses trained ASR model and web-application, available to translate audio files to a text.

Research methods. Machine learning methods, automatic speech recognition methods and technologies, natural language processing methods, mathematical statistics and probability theory.

Scientific novelty of the research. The thesis proposes scientific and practical novelties, which were applied to practical tasks, especially for improving end-to-end automatic speech recognition systems for agglutinative language - focusing Kazakh language and which can easily be applied to other languages. Moreover, contributions were made to the increase of training data size for Kazakh language. The main positive results, obtained during the research are listed below:

1) Was developed data corpus for agglutinative languages.
2) Were developed effective models for recognition of low-resource agglutinative languages from Turkic family: transfer, multilingual, extended language model.
3) System for automatic speech recognition for agglutinative languages.

Theoretical and practical significance of the research. Theoretical importance of the research is that, it proposes optimal continuous attention layers in conformer encoder for agglutinative languages, proves the possibility of improving ASR performance improving only the language model with external “Big Text”, and shows the possibility of improving performance for all languages included in multilingual training for languages from one family group. The possibility of applying all mentioned theoretical statements to train ASR for agglutinative languages of Turkic family shows the practical significance of the current thesis. Moreover, text processing algorithms can be applied to wide range of text processing tasks. Audio-text pair data, collected during research, can be used in different speech processing tasks.

Statements to be defended. Next statements are proposed to the be defended:

1) Dataset for agglutinative languages was developed.
2) Methods of improving ASR for low-resource agglutinative languages were proposed.
3) ASR system for agglutinative languages was developed.

Reliability degree and approbation of the results. Researches and their results related to the thesis topic were presented and discussed in different conferences and seminars and some of them were published. Moreover, the author was awarded with certificates as a seminar speaker, for the best presentation:

1) O. Mamyrbayev, D. Oralbekova, A. Kydyrbekova, T. Turdalykyzy and A. Bekarystankyzy, "End-to-End Model Based on RNN-T for Kazakh Speech Recognition," 3rd International Conference on Computer Communication and the Internet (ICCCI) (25-27 June 2021 y., Tokyo).
2) Certificate to the seminar speaker on the topic “Improved Speech Recognition for Agglutinative languages”, Coimbra Institute of Engineering (ISEC), (21 April 2023 y., Coimbra, Portugal).
3) Certificate for the best presentation speech, “Improve Automatic Speech Recognition for Kazakh Language using Extended Language Model”, “ACeSYRI Young Researchers School” (5-10 June 2023 y., Almaty, Kazakhstan).
4) A. Bekarystankyzy, O. Mamyrbayev, “IMPROVE AUTOMATIC SPEECH RECOGNITION FOR KAZAKH LANGUAGE USING EXTENDED LANGUAGE MODEL”, 21 st scientific conference, (20-21 April 2023y., Riga, Latvia).
5) Automatic Speech Recognition Improvement for Kazakh Language with Enhanced Language Model // Recent Challenges in Intelligent Information and Database systems. ACIIDS 2023. Part of The Communications in Computer and Information Science book series. – 2023. - Vol. 1, - P.538-545 (Springer, Cham).

Personal contribution of the researcher. PhD candidate independently performed and solved the tasks of the PhD thesis. The author designed and implemented end-to-end models for Kazakh and Agglutinative languages. Made own contribution in expanding data corpus for Kazakh language. Designed and performed experimental tests and assessments of the models, both existing and improved models.

The connection of the dissertation topic with the plans of research work. Research works under the research topic were conducted within the grant projects: “Development of an end-to-end automatic speech recognition system for agglutinative languages” (2020-2022, governmental registration number:

0120PK00344) in the Institute of information and computational technologies SC MHES RK.

Main results of the dissertation research. There were four papers published under the research topic, one of which is published in a periodical journal with non-zero impact-factor and indexed by databases Scopus and Web of Science, 3 papers published in the journals recommended by the Control Committee in the sphere of education and science of MHES RK.

1) M. Orken, A. Keylan, O. Dina, B. Akbayan and Z. Bagashar. (2022). Identifying the influence of transfer learning method in developing an end-to-end automatic speech recognition system with a low data level // Eastern-European Journal of Enterprise Technologies. - Vol. 1, № 115. -P. 84–92 // https://doi.org/10.15587/1729-4061.2022.252801(Scopus, percentile 34);
2) Bekarystankyzy A. and Mamyrbayev O. (2023). INTEGRATED AUTOMATIC SPEECH RECOGNITION SYSTEM FOR AGGLUTINATIVE LANGUAGES // News of the National academy of sciences of the republic of Kazakhstan. - 2023. - Vol. 1, № 345. -P. 37-49 // https://doi.org/10.32014/2022.2518-1726.167
3) Bekarystankyzy A., Mamyrbayev O., Oralbekova D., Zhumazhanov B. (2023). Transfer learning for an integrated low-data automatic speech recognition system // Scientific and technical journal "Bulletin of the Almaty University of Power Engineering and Telecommunications". - 2023, -Vol. 1, №. 60. -P. 185-198 // https://doi.org/10.51775/2790-0886_2023_60_1_185
4) Bekarystankyzy A. and Mamyrbayev O. (2023). End-to-end speech recognition systems for agglutinative languages // Scientific Journal of Astana IT University. - 2023. -Vol. 13. -P. 86-92 // DOI: 10.37943/13IMII7575
5) Author’s certificate "Software Product UniCodeKaz" No 38545 from 21.08.2023, Authors:Bekarystankyzy A.
6) Author’s certificate "System of transcribing audio files to text" No 38833 from 31.08.2023, Authors: Bekarystankyzy A., Mamyrbayev O., Duisenkhan B.

Structure and size of the thesis. Dissertation thesis consists of the Introduction, 4 sections, conclusion, bibliography from 163 references, and 4 appendixes. Work is presented in 107 pages and contains 38 figures, 16 tables and 68 equations.

Brief description of the thesis

Introduction presents information about relevance of the research topic, purpose, research questions and objectives of the thesis, scientific novelty, theoretical and practical significance of the research and research methods used in the thesis.

The **first** chapter presents a comprehensive review of related work, both in agglutinative and non-agglutinative languages. Initial discussion in this chapter is dedicated to the methodology of studying state-of-the-art cases. Further, first ASR examples, introduction of neural networks in speech processing area, modern models for ASR are observed. The last two subsections of this chapter were about studies and achievements for agglutinative languages and Kazakh language.

The **second** chapter dedicated to theoretical foundations about speech recognition reviews the most important concepts about agglutinative languages, natural language processing, automatic speech recognition, performance metrics and speech recognition models, starting from mathematical and acoustic modelling, continuing it with the discussion of basic approaches, like hidden Markov models, Gaussian mixture models, hidden Markov Models/artificial neural networks, end-to-end models and the types of network units for sequence models, like RNN, GRU and LSTM. The last subsection describes the state-of-the art architectures based on attention mechanism and connectionist temporal classification: transformer, conformer and branchformer.

The **third** chapter is about experiments and their results. This chapter started from data collection. During the data collection process data in noised situations was collected, like phone conversations, different meeting records and news channels. This process included several stages:

1. 283 hours of data, previously collected in the laboratory of the Institute of Information and Computing Technologies.
2. Data collected for the 2022 year and marked in the laboratory of the Institute of Information and Computing Technologies. This audio data contains phone conversations, audio recordings of zoom meetings, news channels: 195 hr 11 min 25 sec (It was expected that in the end we would have 478 hours of data, but after removing duplicates, only 407 hours remained. It is very important to exclude duplicates in order to preserve the quality of the resulting ASR model).
3. Writing a script for collecting data with different encodings into one file with UTF-8 encoding (UTF-8, UTF-16, rk1048).

Second subsection of third chapter is dedicated to experiments with multitraining example for languages from Turkic family with Cyrillic scripts, like Kazakh, Kyrgyz, Tatar, Bashkir and Saha. This approach was taken into account, because most of existing combining methods of datasets of different languages does not take into account relations of languages to each other. The basic idea of this research is to study the impact of combining languages from Turkic language family, with similar scripts. Common word and sentence formation rules of the selected languages with similar scripts allow to get a working model for each of languages included in the experiments. The contributions of this research are:

1. ASR development method for critically low-resource languages.
2. Improving ASR performance for languages from one language family.

Third subsection of third chapter is about using language model trained on "Big text" in the decoder of end-to-end ASR. Because it is evident, that the rich set of parameters which describes words' relationships and can help to improve the recognition performance of Automatic Speech Recognition Systems (ASR). As a result, inclusion of LM, trained on the big raw text decreased both types of error rates: WER and CER. Especially it has a significant impact on WER. Sequential RNNLM of big text decreased WER by 5%, transformer LM of the same text data decreased WER by 7.2%. Results of experiments showed that Transformer LM is more effective as it can decrease perplexity and increase the number of trainable parameters.

Fourth subsection of third chapter is about transfer learning between representatives of Turkic languages, because most of these languages are low-resource. Also the performance of ASR systems built by end-to-end methods depends on the size of data to be used in training process. Introducing the complex of computational layers can result to huge number of parameters which cannot be reached by training low-resource languages. In order to avoid parameter leakage problem for each language the transfer learning method was applied for languages which belong to Turkic language family: Kazakh and Azerbaijani languages.

Transfer learning is the approach which adapts the models which was trained on one data to another collection of data for training. This research showed next three improvements:

1. The representations taken in the result of training for one language (Kazakh) reduces training time for other language against the training from scratch.
2. Transfer learning allow to use less data for another language for evaluation
3. GPU memory usage decreases because transfer learning does not need the support for gradients of all layers.

The fourth chapter discusses results of all experiments included in the dissertation thesis and their results, moreover these results were compared with the results of previous studies similar to obtained tasks.

The **conclusion** summarizes main results of studies, included in dissertation thesis and includes information about future work which will be provided afterwards.