The Developing of the System for Automatic Audio to Text Conversion

Oleh Basystiuk, Natalya Shakhovska, Violetta Bilynska, Oleksij Syvokon, Oleksii Shamuratov, Volodymyr Kuchkovskiy

[a] *Lviv Polytechnic National University, 12 Bandera str., Lviv, 79013, Ukraine*

Abstract

The paper describes possibilities, which are provided by open APIs, and how to use them for creating unified interfaces which is based on recurrent neural network. In last decade AI technologies became widespread and easy to implement and use. One of the most perspective technology in the AI field is speech recognition as part of natural language processing. New speech recognition technologies and methods will become a central part of future life because they save a lot of communication time, replacing common texting with voice/audio. In addition, this paper explores the advantages and disadvantages of well-known chatbots. The method of their improvement is built. The algorithms sequence-to-sequense based on recurrent neural network is used. The time complexity of proposed algorithm is compared with existed one. Scientific novelty of the obtained results is the method for converting audio signals into text based on a sequential ensemble of recurrent encoding and decoding networks. The practical significance is the modified existing chatbot system for converting audio signals into text.

Keywords[1]

machine translation, deep learning, recurrent neural networks, performance, Keras, PyTorch, sequence-to-sequence

Introduction

Today, the creation of programs simulating human communication remains relevant. The simplest model of communication is the database of questions and answers to them [1]. In this case, there is the problem of describing the knowledge base and the implementation of the interpreter program.

The markup language of the knowledge base can include question patterns and corresponding response patterns. Chatbot can perform additional functions. Most of these functions have an implementation on the Internet and are available as an external API.

The aim of the paper is the recognition of user's gender for making chatbot's answer more likeness that is human. The algorithm for analyzing and parsing the user's text for automatically generating the response of the chatbot is developed.

The object of research is the process of automated conversion of audio signals into text. Result of the paper is the developed software product in the form of a chatbot, which converts the received audio message into text and returns it in the format of a text message.

This algorithm takes into account the topics of correspondence and morphology of the text. The algorithm's work will be based on prefix function and hash function. To add, a comparison of the developed algorithm with the existing ones will be made. This research will describe the way in which was created an interface for Telegram chatbot, whose main aim is to translate audio messages into text.

Scientific novelty of the obtained results is the method for converting audio signals into text based on a sequential ensemble of recurrent encoding and decoding networks.

The practical significance is the modified existing chatbot system for converting audio signals into text.

IT&AS'2021: Symposium on Information Technologies & Applied Sciences, March 5–6, 2021, Bratislava, Slovak Republic

EMAIL: obasystiuk@gmail.com (O. Basystiuk); nataliya.b.shakhovska@lpnu.ua (N. Shakhovska); bilynskavio@gmail.com (V. Bilynska), oleksiy.syvokon@grammarly.com (O.Syvokon), jonburned@gmail.com (O.Shamuratov), forgta.team@gmail.com (V.Kuchkivskiy)
(ORCID: 0000-0003-0064-6584(O. Basystiuk); 0000-0002-6875-8534 (N. Shakhovska); 0000-0002-1609-9813 (V. Bilynska), 0000-0003-4915-6762 (O. Syvokon), 0000-0003-1913-5362 (O.Shamuratov), 0000-0002-7941-6329 (V.Kuchkivskiy)

Literature review

Systems of machine translation of unstructured data from one language to another are modeling work of a human translator. Their productivity depends on their ability to comprehend the language grammar rules. In the translation, the main units are not single words, but phrases or phraseological units expressing various concepts. Only by using them, more complex ideas can be expressed via the translated text.

The main feature of machine translation is the different length for input and output. To be able to work with different input and output length, you need to use a recurrent neural network [1-3]. Initially, the work of computer programs for translation is to replace words or phrases from one language with words or phrases from another. However, then there is a problem that such a replacement cannot provide a quality translation of the text because it requires the definition and recognition of words and whole phrases from the original language.

Machine translation basically performs the replacement of one language words to another language words, but usually, the translation made in this way is relevant, because in order to fully convey the meaning of the sentence and find the most suitable analog in the "target" language - it is often necessary to translate the whole phrase in general.

Solving this problem with statistical and neural translation systems is a rapidly growing field that leads to improved translation, upgrade differences in linguistic typology, better handling differences in linguistic typology, the translation of idioms, and the identification of anomalies.

Modern machine translation software has the function of changing the settings for the domain - industry or professional activity, for example, meteorological reports. By limiting the scope of permissible substitutions/substitutions, we are able to obtain a better translation result [2 – 4]. This method is especially effective in areas where the formal or template-style language is used.

This means that machine translation is more efficient in government and legal documents, rather than translation any less standardized texts. Improving the quality of the final result can also be achieved through human intervention: for example, some systems will be able to provide a more accurate translation if the user will indicate in advance the correct translation of some words in the text.

There are two fundamentally different approaches to the construction of machine translation algorithms: rule-based and statistical-based. The first approach is traditional and is used by most machine translation system developers.

Rule-based MT (RBMT), “Classic Approach” (MT) is a machine translation system based on linguistic information from unilingual, bilingual, or multilingual dictionaries and grammar rules, source language and target language [5].

The system covers the basic semantic, morphological, and syntactic patterns of each language. Accordingly, in order to make a translation, the system must make a preliminary morphological, syntactic, and semantic analysis of the text, and only after that it generates a sentence.

The biggest disadvantage of RB-translation is that in order for a program to perform a correct translation, its database must contain all spelling variations of word entry, and for all cases of ambiguity, lexical selection rules must be written [6]. In itself, adaptation to new domains is not such a complicated process, because the basics of grammar for all domains are the same, and the settings of the areas of user activity are limited only by the correction of lexical selection.

Thus, such a machine translation system is the classical method of its implementation, it allows to obtain a better result than the statistical method, but synthesizes translation more slowly. Statistical machine translation is a type of text-based machine translation that is more effective in working with bigger volumes of language pairs. Language pairs - text data that contain sentences in one language and the corresponding sentences in another. Thus, statistical machine translation has a feature of self-learning. The more language pairs available to the program and the more accurately they correspond to each other, the better the result of statistical machine translation.

The term "statistical machine translation" refers to a general approach to solve the problem of translation, which is based on finding the most probable translation of a sentence using data obtained from a bilingual set of texts [7]. An example of a bilingual set of texts is parliamentary reports, which are minutes of debates in parliament. Bilingual parliamentary reports are issued in Canada, Hong Kong, and other countries; official documents of the European Economic Community are issued in 11 languages, and the United Nations publishes documents in several languages. As a result, these materials are highly useful resources for statistical machine translation.

This system is based on the statistical calculation of the probability of coincidences. To translate, the program must have access to hundreds of millions of documents that have been translated by humans in advance. Such documents serve as templates for the system, on the basis of which it translates. The more documents, the higher the probability of better translation.

At the beginning of its existence, in 2006, Google Translate was based on the statistical method of machine translation, and its translation was of very low quality and was considered one of the worst translation options that can be done by an online translator [8]. Today, Google uses the "neural" method of machine translation (MT) and is in serious competition with commercial enterprises, whose products are not free. Neural network approach is based on the method of deep learning.

Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader group of machine learning methods based on the interpretation of learning outcomes, as opposed to algorithms for specific tasks. Training can be supervised or unsupervised. In recent years, Hybrid machine translation (HMT) has become increasingly popular, and the main technology of implementing HMT become RNN.

Recurrent neural network (RNN) - is a class of artificial neural network, which has connections between nodes. In this case, the connection refers to the connection from the more distant node to the less distant node. The presence of connections allows RNN to memorize and reproduce the entire sequence of reactions to one stimulus. From the programming point of view in such networks there is an analog of the cyclic execution, and from the systems point of view - such networks are equivalent to a finite-state machine. RNNs, are generally used to handle the sequence of words in the processing of natural language [9]. Usually for word sequence processing using the Hidden Markov Model (HMM) and the N-program language model.

Hidden Markov Model (HMM) is the statistical model that simulates the work of a process similar to a Markov process with unknown parameters and the task is to guess unknown parameters on the basis of the observed ones [10]. The obtained parameters can be used in further analysis in a normal Markov model, the state is known to the observer, so the probability of transitions is one parameter.

In NMM it is possible to observe only variables that are affected by this state. Each state has a probabilistic distribution among all possible output values. Therefore, the sequence of words generated by NMM gives information about the sequence of states. The NMM can be considered as the easiest Bayesian network.

Bayesian network - the graphical model in the form of a directed acyclic graph, each vertex of which corresponds to a random variable, and the arcs of the graph encode the relations of conditional independence between these variables [11]. The vertices can represent variables of any type, be weighted parameters, hidden variables, or hypotheses.

There are effective methods that are used to calculate and study Bayesian networks. For conducting a probabilistic output in Bayesian networks, both precise and approximate algorithms are used.

The papers [12, 13] present the Neural-Like Structures based on Geometric Data Transformations. The main advantages of the proposed method are the following: not iterative training process, the high performance in training process, which creates conditions for solution of large-dimension tasks. This approach allows the time complexity reduction, but the number of model's parameters is the same.

The paper [14] propose GMDH-neuro-fuzzy system with small number of hyperparameters but with huge time complexity.

The papers [15 – 20] describe machine-learning algorithms for different signals processing. However, the nature of natural text is not analyzed.

Materials and Methods

At a high-level representation of a recurrent neural network (RNN), shown on Figure 1, it's processes data sequences, such as sentences, one element at a time while retaining a memory (called a state) of what has come previously in the sequence. Recurrent means the output at the current time step becomes the input to the next time step. At each element of the sequence, the model considers not only the current input, but what it remembers about the preceding elements. The most popular cell approach nowadays is the LSTM (Long Short-Term Memory) which maintains a cell state as well as a carry for ensuring that the signal (information in the form of a gradient) is not lost as the sequence is processed. At each time step the LSTM considers the current word, the carry, and the cell state.

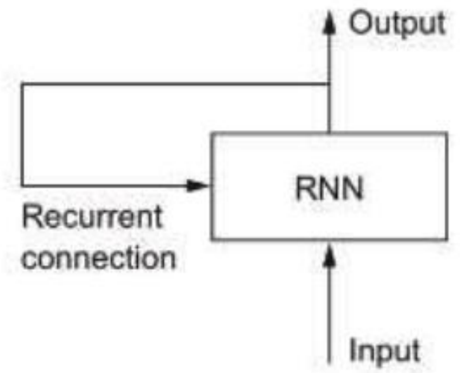


Fig. 1. Recurrent network loop.

The basic idea of an RNN is to use recursion to form the fixed dimension vector from the input sequence of symbols. Assume that in step t vector is h_{t-1} which is the history of all previous words. RNN will calculate new vector ht (its internal state), which combines all previous words ($\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_{t-1}}$) and new character xt using:

$$h_t = \varphi_\theta(x_t\,,\, h_{t-1}).$$

In this equation, the following parameters are present: $\boldsymbol{\varphi\theta}$- function, parameterized with θ, which receive a new word input $\boldsymbol{x_t}$ and words history $\boldsymbol{h_{t-1}}$till (**t** - 1) - N word. First, we can assume that $\boldsymbol{h_0}$ is zero vector. The recurrent activation function φ is usually implemented as an affine transformation, followed by non-linear function:

$$h_t = tanh(Wx_t + Uh_{t-1} + b).$$

In this equation, the following parameters are present: input weight matrix W, recurrence weight matrix U and bias vector b. Note, that this is not the only one variant. There is wide scope for developing new recurring activation functions. More detailed about the work of the method for text translation based on neural networks. The idea of this algorithm is, in fact, simple and consists of the following steps:

1. Encoding the input data of language A into the data set;
2. Decoding the data set in language B.

Let's look at an algorithm for encoding unstructured data on an example text sentence: "Example of neural network" (Figure 2).

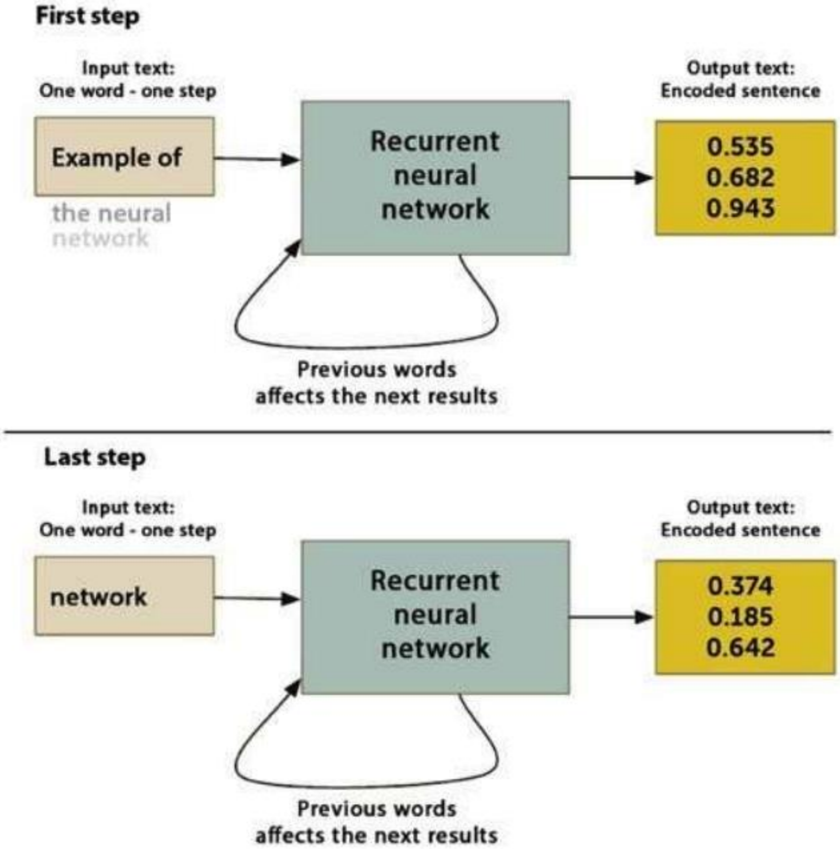


Fig. 2. Visualization of input unstructured data encoding

After performing such a simple operation, we obtain the encoded unstructured data, for example text, that looks like a numerical data set. At the initial stage of training, these numbers are random and generated by the algorithm also accidentally. Next passing of the text that has already encoded, RNN will be evaluated to the same numerical data set. The algorithm of decoding of the unstructured data works like encoding, only in the reverse - the input receives a numerical data set and outputs the probable text that corresponds to this data.

Once we understand the essence of encoding and decoding of the unstructured data, let's move to the very essence of our task - machine translation and its general algorithm. To do this, we just have to combine these two RNNs - for encoding and decoding - and get the following result: Thus, we obtain the general way of transforming the sequence of Ukrainian words into an equivalent sequence of English words, this is the so-called, sequential method of language translation Sequence-to-Sequence. The main pros of the method the following:

- The proposed approach is limited on the training data set amount and the computing power that you can allocate to the translation. Researchers of machine learning have invented this method only a few years ago, but such systems are already working better than the machine translation statistical systems, which was developing through last 20 years;
- The system does not depend on knowledge of any rules of the language. The algorithm itself defines these rules and constantly adapted. The lower-level titles remain unnumbered; They are in the form of run headers.

Results

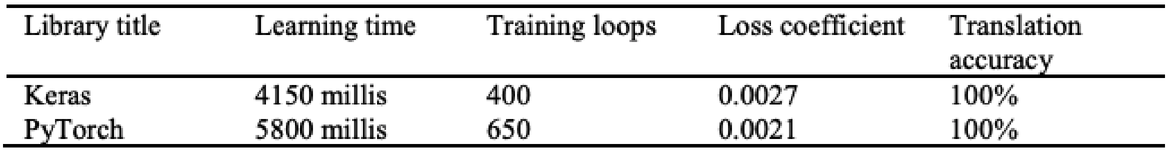
Let’s conduct more information about our dataset and how we will collect that data. First and the most obvious way to collect data is to use open-source datasets, but this way of mining data is not so suitable, in case data will be noisy and will require a lot of economic resources to get from this data high accuracy results in any unique case. Another case is to create own dataset, this is a better way to create personalized solutions for any type of data. The main way to evaluate how noisy is current dataset is to calculate entropy.

$$H(x) = E[\log 1\ p(X)\ ] \leq \log E[\ 1\ p(X)\ ] = \log N$$

As you can see, the training data set consists of 10 phrases, that are widely used in open data sources related to legal cases, we will use that data to train and test our models, based on RNN approach, build on different ML libraries. After that will evaluate the speed and accuracy of the models.

Let’s conduct experiments based on two machine learning libraries written in Python - PyTorch and Keras. The basis of the algorithm is the method of sequential learning.

Table 1. Comparison of Keras and PyTorch libraries results



Library title	Learning time	Training loops	Loss coefficient	Translation accuracy
Keras	4150 millis	400	0.0027	100%
PyTorch	5800 millis	650	0.0021	100%

Let's look at these data in more detail:

- Learning time. The value that shows the model's training time. Mainly depend on the environment where the script was run. Environment mean the current PC specifications; processor computing power and it upload by other processes.
- Training loops. The value that shows training cycles of the model. We give it ourselves.
- Loss coefficient. The value that shows the accuracy of the trained model. It is a measure of how good your model is.
- Translation accuracy. The value that shows in percentage term value of correct translation sentences.

So, the model build on the Keras library was more effective than the PyTorch model, the comparison based on the training time, training loops and error rate. Because of the small training data set, both algorithms show the maximum translation accuracy. In the case of increasing of training data set amount, models will provide completely.

To create a chatbot system for converting audio signals into text, it is necessary to develop an intermediate programming interface (API) for interaction with third-party systems, evaluate and select the optimal technology for backend and interface system, a set of methods and tools for learning, and choose tools for creating a visual design of web pages.

Based on the analysis of content styling technologies, Bootstrap 4, the ngx-bootstrap directive, was chosen because it contains a set of proprietary components, which will greatly simplify use and configuration. In addition, this technology provides detailed documentation and real-life examples (you can run and view the result in real time).

In the process of analyzing the database technology, it was decided to choose MySQL because it is easy to use, configure, design and does not require many resources. It will be inferior to alternatives such as PostgreSQL or Oracle, but the latter requires a paid license, and with the former MySQL can be on par with data processing speeds of up to several thousand, after optimization.

In the process of analyzing the libraries used in the development of the machine learning system, it was decided to choose Keras because it is more efficient, easier to use, has templates for creating and learning neural networks and requires fewer resources. Of course, it is inferior to alternatives like TensorFlow, but this solution is more complex and requires additional costs to configure the system before starting work, as well as additional support after implementation.

The advantages and disadvantages of several approaches, such as rulebased, statistical, and neural network-based are described. Considering all the factors, the most relevant way of organization and software approach for creating methods for analyzing open data in legal cases [5].

Moreover, overviewed design and software approach of the two systems for numbering unstructured data based on different ML frameworks was chosen. For example, this solution will be suitable for translating sentences from one language to another. In the case of an RNN-based language translation approach, the most popular ML libraries are Keras and PyTorch.

System deployment diagram is given in Fig. 3

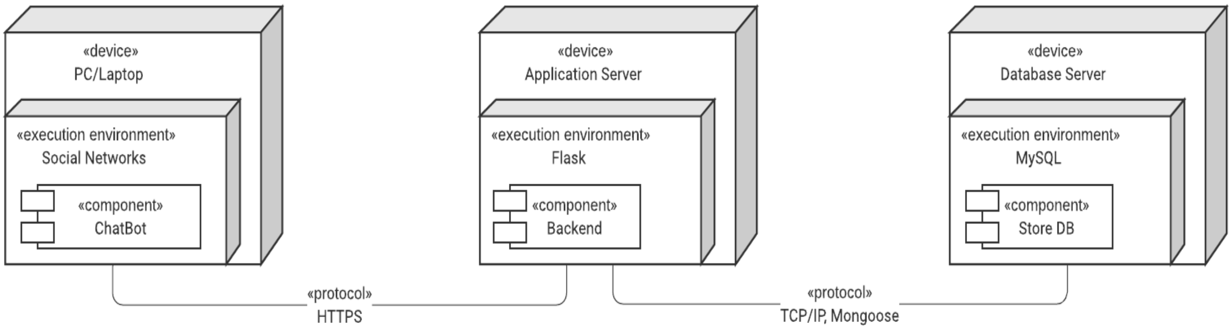
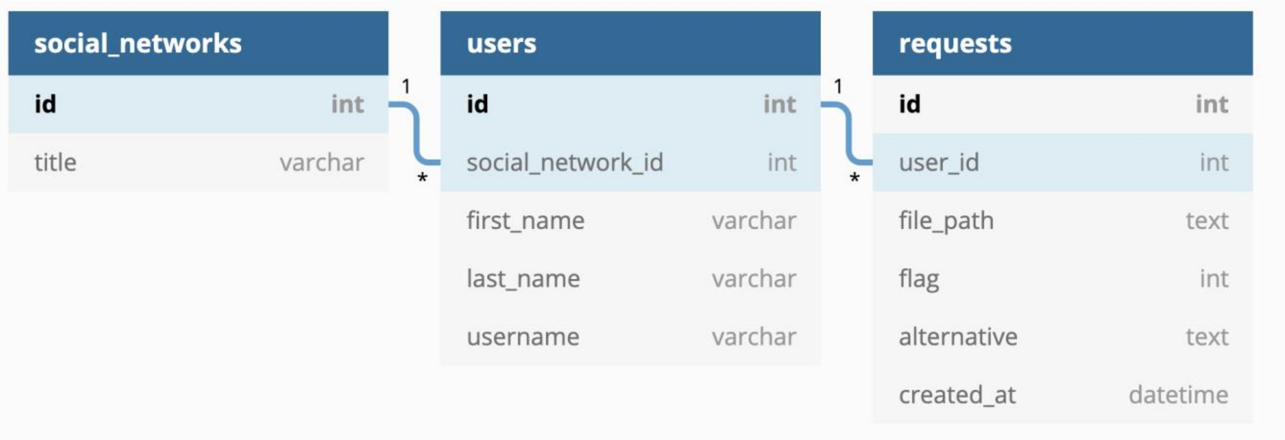


Fig. 3. System deployment diagram

The schema of developed database is given on Figure 4.



social_networks	
id	int
title	varchar

users	
id	int
social_network_id	int
first_name	varchar
last_name	varchar
username	varchar

requests	
id	int
user_id	int
file_path	text
flag	int
alternative	text
created_at	datetime

Fig. 4. The database schema

The main functions of this system:

- receive an incoming user message and process it;
- the received message is checked for audio;
- the received audio file is processed and converted into text;

- send the result in the form of a response to the user

Conclusion

The result of this work is a system for automatic conversion of audio into text (called Harry Bot) with improved, in accordance with analog systems, performance, ease of use and implemented on the basis of modern technologies of artificial intelligence and machine learning by self-learning and continuous improvement the results of the transformation

RNN, like other classes of neural networks, are developing so fast that it's increasingly difficult to track new, more interesting, and more sophisticated models for solving more complex and complicated tasks. These sequential methods of teaching neural networks can be used in other areas, not only in machine translation. Simple examples are models that could make verbal descriptions of the image, recognize the voice and maintain the conversation. In our opinion, the development of RNN will lead to the emergence of smart assistants that can recognize the owner's voice and correctly perceive the task.

At the moment RNNs are the most frequently used in machine translation and we think this field will be also upgraded in the nearest future. According to the results of the experiment, the model based on Keras library is more efficient for the current training data set. Note, that the research results may be considered relevant only for small data sets and there will be changes in translation quality and training time after increasing the training data set amount. Next phase of this research may consist of model training in large data volumes with analyzing and comparing the quality and speed of its work.

References

[1] Boyko, N., Basystiuk, O.: Comparison Of Machine Learning Libraries Performance Used For Machine Translation Based On Recurrent Neural Networks, 2018 IEEE Ukraine Student, Young Professional and Women in Engineering Congress (UKRSYW), pp.78-82, Kyiv, Ukraine (2018)

[2] Shakhovska, N., Basystiuk, O., & Shakhovska, K. (2019). Development of the Speech-to-Text Chatbot Interface Based on Google API. In MoMLeT (pp. 212-221).

[3] Goodfellow I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. – Berkley : The MIT Press, 2016. – 775c.

[4] Boyko N., Pylypiv O., Peleshchak Y., Kryvenchuk Y., Campos J.: Automated document analysis for quick personal health record creation. 2nd International Workshop on Informatics and Data-Driven Medicine. IDDM 2019. Lviv. p. 208-221. (2019)

[5] Inurrieta, U., Aduriz, I., de Ilarraza, A. D., Labaka, G., & Sarasola, K. (2017, April). Rule-based translation of Spanish Verb-Noun combinations into Basque. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) (pp. 149-154).

[6] Cheng, Z., Tisi, M., & Douence, R. (2020). CoqTL: a Coq DSL for rule-based model transformation. Software and Systems Modeling, 19(2), 425-439.

[7] Chen, K., Zhao, T., Yang, M., Liu, L., Tamura, A., Wang, R., ... & Sumita, E. (2017). A neural approach to source dependence based context model for statistical machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(2), 266-280.

[8] De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. Political Analysis, 26(4), 417-430.

[9] Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., ... & Zhang, W. (2019, December). A comparative study on transformer vs rnn in speech applications. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 449-456). IEEE.

[10] Xue, C. (2018, August). A novel english speech recognition approach based on hidden Markov model. In 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS) (pp. 1-4). IEEE.

[11] Chaturvedi, I., Ragusa, E., Gastaldo, P., Zunino, R., & Cambria, E. (2018). Bayesian network based extreme learning machine for subjectivity detection. Journal of The Franklin Institute, 355(4), 1780-1797.

[12] Tkachenko R., Izonin I. (2019) Model and Principles for the Implementation of Neural-Like Structures Based on Geometric Data Transformations. In: Hu Z., Petoukhov S., Dychka I., He M. (eds) Advances in Computer Science for Engineering and Education. ICCSEEA 2018. Advances in Intelligent Systems and Computing, vol 754. Springer, Cham.

[13] Izonin, I., Tkachenko, R., Vitynskyi, P., Zub, K., Tkachenko, P., & Dronyuk, I. (2020, November). Stacking-based GRNN-SGTM Ensemble Model for Prediction Tasks. In 2020 International Conference on Decision Aid Sciences and Application (DASA) (pp. 326-330). IEEE.

[14] Zhao, C., Ni, B., Zhang, J., Zhao, Q., Zhang, W., & Tian, Q. (2019). Variational convolutional neural network pruning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2780-2789).

[15] Postma, B. N., Poirier-Quinot, D., Meyer, J., & Katz, B. F. (2016). Virtual reality performance auralization in a calibrated model of Notre-Dame Cathedral. Euroregio, 6, 1-10.

[16] Wawrzonowski, M., Daszuta, M., Szajerman, D., & Napieralski, P. (2017, September). Mobile devices' GPUs in cloth dynamics simulation. In 2017 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 1283-1290). IEEE.

[17] Napieralski, P., Juszczak, E. N., & Zeroukhi, Y. (2016). Nonuniform distribution of conductivity resulting from the stress exerted on a stranded cable during the manufacturing process. IEEE Transactions on Industry Applications, 52(5), 3886-3892.

[18] Fornalczyk, K., & Wojciechowski, A. (2017, September). Robust face model based approach to head pose estimation. In 2017 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 1291-1295). IEEE.

[19] Walczak, J., & Wojciechowski, A. (2016). Improved gender classification using discrete wavelet transform and hybrid support vector machine. Machine Graphics & Vision, 25(1/4), 27-34.

[20] Fedushko S., Ustyianovych T. (2021) Operational Intelligence Software Concepts for Continuous Healthcare Monitoring and Consolidated Data Storage Ecosystem. In: Hu Z., Petoukhov S., Dychka I., He M. (eds) Advances in Computer Science for Engineering and Education III. ICCSEEA 2020. Advances in Intelligent Systems and Computing, vol 1247. Springer, Cham. pp. 545-557. https://doi.org/10.1007/978-3-030-55506-1_49

[21] M. Zubair Khan, “Hybrid Ensemble Learning Technique for Software Defect Prediction,” IJMECS, vol. 12, no. 1, pp. 1–10, Feb. 2020, doi: 10.5815/ijmecs.2020.01.01.

[22] W. Yasya, “Rural Empowerment through Education: Case Study of a Learning Community Telecentre in Indonesia,” p. 15, 2020.

[23] Fedushko S., Syerov Yu., Tesak O., Onyshchuk O., Melnykova N. (2020) Advisory and Accounting Tool for Safe and Economically Optimal Choice of Online Self-Education Services. Proceedings of the International Workshop on Conflict Management in Global Information Networks (CMiGIN 2019), Lviv, Ukraine, November 29, 2019. CEUR-WS.org, Vol-2588. pp. 290-300. http://ceur-ws.org/Vol-2588/paper24.pdf

[24] Wojciechowski, A., & Fornalczyk, K. (2014, September). Exponentially smoothed interactive gaze tracking method. In International Conference on Computer Vision and Graphics (pp. 645-652). Springer, Cham.