Development of Technology for Summarization of Kazakh Text

Talgat Zhabayev, Ualsher Tukeyev
Department of Information Systems
Al-Farabi Kazakh National University
Almaty, Kazakhstan

***Abstract*—This paper presents the solution to the problem of summarizing Kazakh texts. The problem of Kazakh text summarization is considered as a sequence of two tasks: extracting the most important sentences of the text and simplifying the received sentences. The task of extracting the most important sentences of the text is solved using the TF-IDF method and the task of simplifying sentences is solved using the neural network technology "Seq2Seq". Problem of using NMT method for simplification of Kazakh was in absence of Kazakh dataset for training. To solve this problem in this work propose use transfer learning method. The use of transfer learning made it possible to use a ready-made model that was trained on a parallel corpus of Simple English Wikipedia and not create a simplification corpus in Kazakh from scratch. For this, a transfer learning technology for simplifying sentences of the Kazakh language has been developed, based on training a neural model for simplifying sentences in the English language. Main scientific contribution of this work is transfer learning technology for the simplification of Kazakh sentences using the parallel corpus of the English language simplification.**

Keywords*—Summarization; text simplification; low-resource language; seq2seq; transfer learning

I. Introduction

Automatic text summarization is the process of shortening text without losing meaning. It can be the text of one or several documents. Summarization has become widespread in recent years in many application areas. Practical applications - data analytics, automatic creation of headlines or short descriptions, news sites, information aggregators. In these tasks, there is a need for automatic annotation of a huge amount of text data, which is inefficient to perform manually; by automating these actions, you can achieve significant time savings.

There are two types of annotation - abstractive and extractive. Extractive summarization - highlights the most important sentences in a text that most fully describe this text. Abstractive summarization is the reduction of a text by paraphrasing the text into a short form. In this case, the final summarization may contain phrases or sentences that did not occur in the original text.

Text simplification is an area of study in computational linguistics that studies methods and techniques for simplifying textual content [1]. In Natural Language Processing it is used as one of the steps in summarization, text parsing [2], text translation, question-answer systems. Simplification is performed by shortening sentences, combining sentences, transforming sentences, paraphrasing. In this work, annotation is considered for the Kazakh language.

To implement transfer learning, we use the second parallel corpus Kazakh - English. A model trained on a large parent corpus of the English language should give a relatively high-quality result of simplification and be based on the quality of the model and not on the knowledge of the language. The relevance of the study is due to the fact that at present, research in the field of annotating the Kazakh language is focused on extractive summarization and little attention is paid to the abstractive method. The main reason of this situation is the absence of the Kazakh corpora for abstractive summarization and the difficulty of creating it. Applying transfer learning in this work, we use sequentially extractive and abstractive summations to obtain a short version of the text.

The scientific contribution of this work is: 1) in the development of a TF-IDF [3] model for texts of the Kazakh language, using the Kazakh language corpus, processing it to obtain frequencies for TF-IDF; 2) in the development of transfer learning technology for the simplification of Kazakh sentences using the parallel corpus of the English language simplification.

The remaining part of this paper is organized as follows. Section II contains an overview of existing papers on summarization, the use of neural networks for simplification. Section III contains the application of the neural network technology "Seq2Seq" to simplify text. The machine translation of the neural network technology "Seq2Seq" is based on the use of parallel data corpuses. After training, the model is able to generate the simplification of new sentences. Section IV describes the implementation of abstractive summarization and sentence extraction using the TF-IDF method. Section V contains a description of training the model and the results in the form of a table. Section VI concludes the overall study.

II. Related Work

Consider research on abstractive summation. At the moment, to implement this type of summarization, the most common method is using neural networks with "sequence to sequence" architectures. Initially, sequence to sequence neural networks were used in neural machine translation. The architecture of a neural network describes the number, types of layers, the number of neurons in them and how the layers are connected to each other. Seq2seq neural networks are used

together with the element of attention [4]. This type of neural networks consists of a decoder and an encoder, which are recurrent neural networks [5]. The use of sequence to sequence architecture for machine translation has been described in many works [6, 7].

Currently, the most used architecture in the summarization of text is the transformer [8]. The difference between the transformer architecture and seq2seq is parallel, not sequential processing of input sentences. Transformer is the so-called attention-based architecture. Simplification model training is distinguished primarily by the training corpus. A parallel corpus for simplification problems is a corpus, the source part of which is the ordinary language sentences, and the target part is the corresponding simplified language sentences. Thus, simplification is a monolingual task for neural machine translation.

The main corpus for simplification is the Simple English Wikipedia corpus [9]. Training on this corpus forms the basis of most of the papers on text simplification. So in [10] this simplification corpus was used to train the seq2seq simplification model.

The model was trained in the OpenNMT system [11]. It is one of the most popular tools for neural machine translation. There are several implementations – original OpenNMT, OpenNMT-Python, OpenNMT-Tensorflow. Similar papers in the field of simplification are [12,13] where a neural transformer model is also used.

There are many works available on the summarization of texts in low-resource languages. Transfer learning is also increasingly used for low-resource languages. For example, in [14, 15], the use of neural networks for abstractive summation together with the transfer of learning is considered.

In [16], the authors describe the creation of a synthetic set of complete sentences for simplification using a pretrained model. Neural networks are also used for extractive summation. In [17], a summarization corpus is used, where source is a set of ordinary sentences, and the target part of the corpus is the summation of the corresponding set of sentences.

In the [18], the authors used centroids and Word's mover distance for extraction summarization in Kazakh language. Many summarization studies look at TF-IDF and data clustering. Also TF-IDF is used in information extraction [19].

When defining sentences for extractive summarization, it is need to get those sentences that together describe the text as much as possible (and there should be no unnecessary, redundant sentences in summarization) [20]. Work [21] describes a similar implementation of summarization using TF-IDF.

Transfer learning methods find application in the case of low resource languages, such as in [22], where the authors used the general parent model and the child model to translate the Tibetan language. Transfer learning is an area of NLP research that focuses on the problem of retaining knowledge that was obtained by training one model and transferring knowledge to another, similar problem [23, 24].

III. Methodology of Summarization of Kazakh Text

The proposed methodology of summarization of Kazakh text includes two steps:

- Extraction of summarize sentences,
- Simplification of extracted sentences.

Below these two parts detailed are considered.

A. Extraction of Sentences

The TF-IDF metric is used to implement extractive summarization.

There are several options for using TF-IDF: 1) Ranking sentences by the value of TF-IDF or sentence centroids to find the most important sentences in the corpus; 2) search for the most similar sentences by semantic proximity; 3) clustering of sentences by the values of TF-IDF or centroids [25].

In our work, we use the centroid ranking method and clustering.

Below is a step-by-step implementation algorithm:

1) We perform preprocessing, which includes removing punctuation marks, apostrophes, dashes and other uninformative elements, tokenizing the text using the sent_tokenize function in order to get an array of text, where each element is a separate sentence.

2) We get term frequency - it is defined as the ratio of the number of times each unique word appears in the sentence to the number of words in the sentence.

3) We get inverse document frequency - a value that shows the significance or informativeness of a word in a sentence, allowing you to ignore words that appear in most sentences, such as prepositions. It is the logarithm of the ratio of the number of sentences to the number of occurrences of a word.

4) The centroid of each sentence is calculated as the ratio of the sum of TF-IDF values to the total number of unique words in the sentence.

5) We combine all the centroids of the sentences into one array, which contains the sentence number and the centroid value. Then we select several sentences with the largest centroid values.

In Fig. 1 shows a graph of the distribution of centroid values for a text in the Kazakh language, which was obtained as a result of text simplification. On the chart, the X-axis is the ordinal number of the sentence, the Y-axis is the values of the centroids of the corresponding sentence. Centroid - a value from zero to 1. On the diagram, we see how the centroid values of sentences are distributed in the corpus and in which part of the corpus the largest centroid values are.

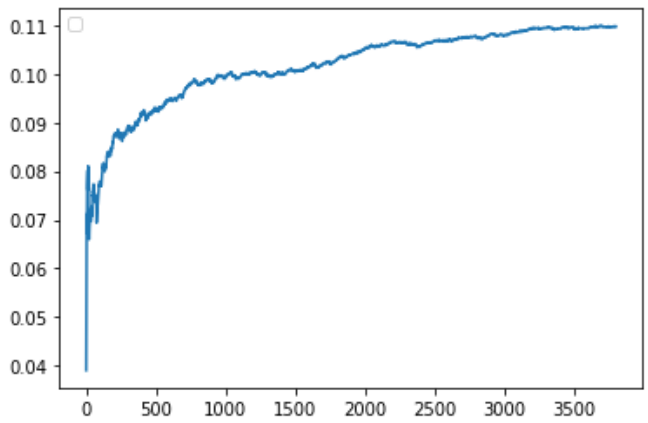

Fig. 1. Distribution of Centroid Values by Sentences of the Kazakh Corpus.

The next step is to analyze the resulting simplified corpus by distributing words into clusters for topic analysis. Table I shows the result of the distribution of the corpus into 6 clusters. For clustering, we use the k-means algorithm [26]. This algorithm allows you to group a set of vectors according to the degree of similarity. In this case, we use centroids as a criterion for the similarity of words. For simplicity, the number of clusters was chosen arbitrarily. As we can see, each cluster contains a set of words that are close in meaning.

TABLE I. WORD CLUSTERING

cluster	sentence
1	эскиздер,байқауына,эскиздерді,жобалау,көзқарас,президент,қызығушылық,қазақстан,иран,бұл
2	республикасының,қазақстан,президенті,президентіне,сауд,сенім,грамоталарын,тапсырды,бар,заңы
3	барлық,қазақстанның,мәселелері,қажет,өсім,ақпарат,салаларында,бар,және,диалог
4	бар,кездесу,жиналыспен,көршілес,қоғамдық,немесе,көптеген,қазақстанда,маңызы,қалада
5	премьер,министр,министрдің,кездесуінде,бұл,министрі,астана,шетелдіктердің,кеңестегі,қала

B. Simplification of Extracted Sentences

In this subsection, we will describe the algorithm for creating a simplification model of the Kazakh text using a method that relates to the transfer learning. The Kazakh language is a language with a small number of parallel corpuses, which makes learning a neural model very difficult. A model for working with the Kazakh language should be trained on a parallel corpus of the Kazakh text. For the simplification of texts in the Kazakh language, there are currently no ready-made simplification parallel corpus. Therefore, to obtain such a corpus, we use the Google translate application with manually edition to translate English parallel simplification corpus to Kazakh parallel simplification corpus.

The proposed methodology of simplification of Kazakh text includes two stages (Fig. 2).

At the first stage, the parent model is trained:

1) First, we define the architecture of the parent model. Before creating the model, it was necessary to choose the architecture of the neural network model that would show the highest score values in the original English corpus. To do this, we train seq2seq and a transformer model on a general corpus (Simple English Wikipedia) and see which architecture has bigger BLEU.

2) The English part of the kaz-eng corpus [27] is translated by the trained model. As a result, we got a simplified text of the English part of the kaz-eng corpus.

Second stage of transfer learning is the training of the child model:

1) The resulting simplified part of the kaz-eng corpus was translated into the Kazakh language, using the public web service of machine translation with the recording of the result in a text file.

2) As a result, a synthetic Kazakh parallel simplification corpus is obtained. The source part of Kazakh parallel simplification corpus is the Kazakh part of the source kaz-eng corpus and the target part is the simplified text of the source English part translated on Kazakh.

3) After that, the training a new neural model on the Kazakh parallel simplification corpus is made.

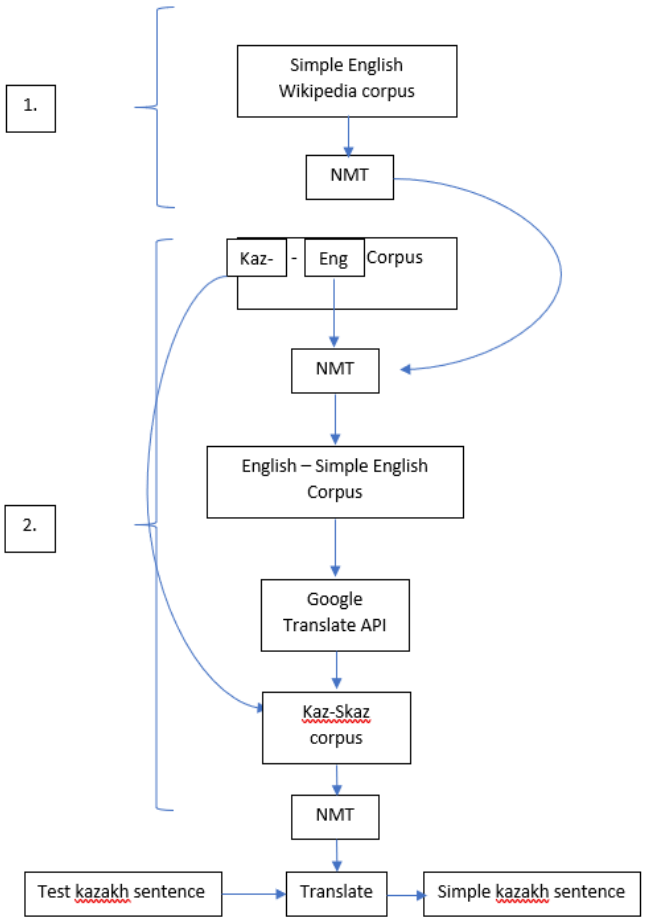

Fig. 2. Transfer Learning Technology for the Simplification of Kazakh Sentences.

The considered method of using the results of model translation to create a simplified part of the corpus belongs to transfer learning methods. These are learning transfer methods that use the generated synthetic data at runtime [28]. A corpus that was created based on the data generated by the model is called synthetic. Also, the creation of a synthetic corpus underlies the method of back-translation [29]. So, in [30] the author uses back-translation to increase the size of the training corpus. In our work, it is possible to use this method to increase the train corpus.

Nevertheless, synthetic data is worse than real data, and when a significant part of the train corpus is synthetic data, the model usually shows worse BLEU results compared to real data [31].

Further, the simplified text in the Kazakh language, allows to define the sentences that convey the essence of the text.

IV. NMT Experiment and Result

This section discusses training the model, obtaining results, and assessing the quality of summarization. Quality assessment is needed to determine how well the model performs, including with text that is very different from the training corpus. To assess the quality of the model, we use the BLEU and SARI metrics.

The BLEU (Bilingual Evaluation Understudy) metric is an assessment of the quality of machine translation from one language to another. The BLEU algorithm compares the number of common words or phrases in predicted sentences with reference sentences. Comparison is performed by counting N-gram matches. The final score for a corpus is the average quality score for all sentences in the corpus [32]. The metric has certain drawbacks in the case of text simplification, since was originally developed for machine translation rather than text simplification [33].

In [34], the SARI (System Output Against References and Input Sentences) metric was presented. This metric can assesses the quality of text simplification based on source, predictions and reference data, correctly taking into account the operations to change the sentence.

When training a model, one of the most important parameters is the number of training epochs and dropout. The Epoch - full cycle through the hull during training, it takes more than one epoch to train the model. Dropout is a technique used in training, which consists in shutting off the outputs of some neurons with a certain probability [35], which avoids overfitting the model. The model works on the basis of a vocabulary that was created during training. The size of the vocabulary affects the performance of the model. Vocabulary size we have set 50 000 words.

The parallel corpus of Simple English Wikipedia contains 284677 lines for training. The model was trained for 20 epochs, until the values of the loss function ceased to decrease significantly. The kaz-eng corpus contains 109 thousand lines, from where 5000 lines are allocated for testing.

To determine the most optimal option, we also applied the model fine-tuning method [36], which refers to inductive transfer learning. It differs from the previous method in the following steps: 1) it is necessary that the model that has been trained on the general corpus is retrained on the domain-specific corpus. To do this, OpenNMT connects the existing model vocabulary to the Kazakh corpus vocabulary; 2) a new savepoint is created in OpenNMT, which uses the new dictionary. The training of the model continues from a new point. Thus, the model trained in English is retrained taking into account the Kazakh language.

Table II shows the scores of BLEU and SARI of neural models depending on the parallel data corpus. These grades are obtained during testing after training the models.

TABLE II. BLEU and SARI Scores for Seq2seq and Transformer Model

№	*Model*	*Simple English Wikipedia BLEU/SARI*	*Kaz-eng BLEU/SARI*	*Kaz-skaz BLEU/SARI*
1	Seq2seq	58.01/52	53/66.18	Not trained
2	Transformer	66.70/66	55/60	7/36
3	Transformer Finetuned	66.70/66	55/60	8/36

Column “Simple English Wikipedia BLEU/SARI” contains the BLEU scores after training the model on the Simple English Wikipedia.

Column “Kaz-eng BLEU/SARI” - BLEU assessment at the stage of translating the English part of the kaz-eng corpus.

Column “Kaz-skaz BLEU/SARI” - BLEU score after training the Kazakh simplication model.

As we can see from the data in Table II for the seq2seq model, the BLEU score has changed from 58 on the Simple English Wikipedia corpus, to 53 on the kaz-eng corpus. The test set is the selected lines from the Simple English Wikipedia train corpus, that is, it is data of a similar subject.The kaz-eng corpus test set for the model was not used for training the model and this corpus was not originally for text simplification.

On the transformer model, the BLEU score is also reduced from 66 to 55. According to the parent data, the model with the transformer architecture has a slight advantage over the seq2seq attention model on the same data. This model is also the parent for fine-tuned Transformer model.

Therefore, the transformer architecture was chosen to create the Kazakh model.

The BLEU score on the resulting Kazakh child model is 7. As we can see from the assessment of the “Transformer Finetuned” model, the assessment increased by 1 and this method of creating the Kazakh model is better.

When translating, the model works with many unfamiliar words, and the meaning of words, depending on the context, may differ. This problem is called domain shift [37]. In other words, a model trained on news data does not work well with data from medicine or another field of science. This is one of the reasons for the low BLEU score on the Kazakh model. Another reason may be an error in training the transformer

model, which affected the quality of the translating, which we will try to fix in the future.

V. Conclusion and Future Work

In this paper, the method for summarizing Kazakh text was considered. Proposed Kazakh text summariszing method based on consequent using of TF-IDF method for extracting summarize sentences and NMT method for simplification of received summarize sentences. Problem of using NMT method for simplification of Kazakh was in absence of Kazakh dataset for training. To solve this problem in our method to propose use transfer learning method. The use of transfer learning made it possible to use a ready-made model that was trained on a parallel corpus of Simple English Wikipedia and not create a simplification corpus in Kazakh from scratch.

In future works, we plane to further improve the model, by increase the volume of the training dataset of the Kazakh corpus. Also we plane investigate using of post-editing NMT technology for increase of Kazakh parallel simplification corpus volume and quality. One of the directions for further research on this area is a method of clustering similar sentences in the train dataset and training a new seq2seq model based on it, as in [38]. Which should improve the performance of the model.

References

[1] Saggion H., "Automatic Text Simplification," Morgan & Claypool Publishers, 2017, p. 2.

[2] Chandrasekar R., Doran C., "Motivations and methods for text simplification," COLING Volume 2: The 16th International Conference on Computational Linguistics, pp. 1041-1042, 1996.

[3] Salton G., Buckley C., "Term-Weighting approaches in Automatic Text Retrieval," Information Processing and Management 24(5), pp. 513-523, 1988.

[4] Bahdanau D., Cho K., Bengio Y., "Neural machine translation by jointly learning to align and translate," 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.

[5] Hochreiter S., Schimidhuber J., "Long short-term memory," Neural Computation:journal, Vol.9, no.8. - pp.1735-1780, 1997.

[6] Sutskever I.,Vinuals O., Q.V. Le., "Sequence to Sequence Learning with Neural Networks," Advances in Neural Information Processing Systems, pp. 3104-3112, 2014.

[7] Cho K., Van Merrienboer B.. Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," EMNLP 2014 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Doha, pp. 1724-1734, 2014.

[8] Vasvani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., "Attention is all you need," Advances in Neural Information Processing Systems 30, pp. 5998-6008, 2017.

[9] Hwang W., Hajishirzi H., Ostendorf M., Wu W., "Aligning Sentences from Standard Wikipedia to Simple Wikipedia," Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies, pp.211-217, 2015.

[10] Nisioi S., Stagner S., Ponzetto S.P., "Exploring Neural Text Simplification Models," ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) , Vancouver, V. 2., pp. 85-91, 2017.

[11] Klein G.,Kim Y., Deng Y., Senellart J., Rush A., "OPENNMT:Opensource toolit for neural machine translation," In Proceedings of ACL 2017, System Demonstrations, Vancouver. Association for Computational Linguistics, pp. 67-72, 2017.

[12] Surya S., Mishra A., "Unsupervised text simplification," ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp. 2058-2068, 2019.

[13] Maruyama T., Yamamoto K., "Extremely low resource text simplification with pre-trained transformer language model," Proceedings of the 2019 International Conference on Asian Language Processing, IALP 2019, pp. 53-58, 2019.

[14] Chowdhury R.R., Nayeem M.T., Mim T.T.,"Unsupervised abstractive summarization of Bengali text documents," EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, pp. 2612-2619.

[15] Quasmi N.H., Zia H.B., Athar A., Raza A.A.,"SimplifyUR:unsupervised lexical text simplification for urdu," LREC 2020 -12th International Conference on Language Resources and Evaluation, Conference Proceedings, 2020 12th International Conference on Language Resources and Evaluation, LREC 2020, Marseille, 164155, pp. 3484 - 3489, May 2020.

[16] Parida S., Motlicek P., "Abstract text summarization: a low resource challenge," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), 2019.

[17] Ambrosio A.P., Tonelli S., Turchi M., Negri M., Di Gangi M.A., "Neural text simplification in low-resource conditions using weak supervision," Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation, 2019.

[18] Seitkali, D., Musabayev, R., "Using centroid keywords and word mover's distance for single document extractive summarization," ACM International Conference Proceeding Series, pp. 149-152, 2019.

[19] Hashemzadeh B., Abdolrazzagh-Nezhad M., "Improving keyword extraction in multilingual texts," International Journal of Electrical and Computer Engineering Open AccessVolume 10, Issue 6, pp. 5909 - 5916, December 2020.

[20] Gholipour Ghalandari D., "Revisiting the Centroid-based Method: A Strong Baseline for Multi-Document Summarization," Proceedings of the Workshop on New Frontiers in Summarization, September 2017.

[21] Christian H., Pramodana Agus M., Suhartono D., "Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)," International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 6, December 2020, pp. 5909~5916.

[22] Zhou M., Secha J., Cai R., "Domain adaptation for tibetan-chinese neural machine translation," ACAI 2020: 2020 3rd International Conference on Algorithms, Computing and Artifical Intelligence December 2020. Article No.: 77 , pp. 1-5, December 2020.

[23] West J., Ventura D., Warnick S., "Spring Research Presentation: A Theoretical Foundation for Inductive Transfer," Brigham Young University, College of Physical and Mathematical Sciences. Archived from the original on 2007-08-01. Retrieved 2007-08-05, 2007.

[24] Malte, Aditya and Pratik Ratadiya, "Evolution of transfer learning in natural language processing," arXiv:1910.07370, 2019.

[25] Radev D., Hongyan J., Stys M., Tam D.. 2004. "Centroid-based summarization of multiple documents," Information Processing and Management 40(6), pp. 919-938, 2004.

[26] Pelleg, D., Moore A., "Accelerating exact k -means algorithms with geometric reasoning," Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '99. San Diego: ACM Press: pp.277–281, 1999.

[27] https://github.com/NLP-KazNU/kaz-parallel-corpora_collect_and_clean.

[28] Pan S.J., Yang Q., "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering (Volume: 22, Issue: 10), pp. 1345-1359, Oct. 2010.

[29] Sennrich R., Haddow B., Birch A., "Edinburgh Neural Machine Translation Systems for WMT 16,"In Proceedings of the First Conference on Machine Translation, pp. 371-376, Berlon, 2016.

[30] Qiang, Jipeng, "Improving Neural Text Simplification Model with Simplified Corpora," arXiv:1810.04428 , 2018.

[31] Wu L., Wang Y., Xia Y., Tao Q., Lai J., Liu T.Y., "Exploiting monolingual data at scale for neural machine translation," Conference

on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pp. 4207-4216, 2019.

[32] Papineni K., Roukos S., Ward T.,Wei-Jing Zhu, ”BLEU: a method for automatic evaluation of machine translation,” Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002, pp. 311-318, 2002.

[33] Sulem E., Abend O., “Bleu is not suitable for the evaluation of text simplification,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pp. 738 - 744, 2020.

[34] Xu W., Napoles C., Pavlick E., Chen Q., “Optimizing Statistical Machine Translation for Text Simplification,” Transactions of the Association for Computational Linguistics, Volume 4. pp 401-415, 2016.

[35] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research 15, pp. 1929-1958, 2014.

[36] Chenhui C., Wang R., “A survey of domain adaptation for neural machine translation,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp.1304-1309, 2018.

[37] Baochen S., Feng J., Saenko K., “Return of frustratingly easy domain adaptation,” 30th AAAI Conference on Artificial Intelligence, AAAI , pp. 2058-2065, 2016.

[38] C. Fan, Yu. Tian, Y. Meng, N. Peng, X. Sun, Fei Wu, Jiwei Li, “Paraphrase Generation as Unsupervised Machine Translation,” arXiv:2109.02950v1, 2021.