Towards Sequence-to-Sequence Neural Model for Croatian Abstractive Summarization

Vlatka Davidović
University of Rijeka
Faculty of Informatics and Digital Technology
Radmile Matejčić 2, 51000 Rijeka
Business Department, Polytechnic of Rijeka
Trpimirova 2/V, 51000 Rijeka
vlatka.davidovic@uniri.hr

Sanda Martinčić Ipšić
University of Rijeka
Faculty of Informatics and Digital Technology
Center for Artificial Intelligence and Cybersecurity
Radmile Matejčić 2, 51000 Rijeka
smarti@uniri.hr

Abstract. *Abstractive text summarization is a natural language processing task of generating a summary from an input text while preserving the meaning of the text. Today, two prevalent deep learning architectures for automatic text summarization are sequence-to-sequence and transformer. The sequence-to-sequence architecture works with sequential data where the order of words in the text is important. The models have reached state-of-the-art performance in English, but the same task is challenging for low-resourced and morphologically rich languages. Moreover, there are few publicly available datasets prepared for training the summarization task. We focus here on Croatian, since the summarization dataset for Croatian is still missing. In this paper, we propose a solution to fill this gap. The first step is creating the summarization dataset in the Croatian language, using Google machine translation from English to Croatian. With the obtained Croatian version of the dataset, we perform initial training of the sequence-to-sequence model with an attention mechanism. The preliminary results of the Croatian abstractive summarization are presented using the evaluation metrics ROUGE and BERTScore.*

Keywords. abstractive summarization, sequence-to-sequence, Croatian

1 Introduction

The main goal of automatic text summarization is to generate a short summary of the input text without human intervention while preserving the main information and meaning (Allahyari et al. 2017). At the same time, the generated summary has to be fluent and similar to the abstract written by humans. The task can be defined as extractive or abstractive text summarization. While extractive text summarization can extract sentences from the original text, abstractive text summarization makes a summary by paraphrasing the main contents of the text with words that may not be in the original text (Widyassari et.al., 2022).

Datasets are an important part of any natural processing task, including automatic summarization, and they are used for training, validation, and testing of the models. Text summarization models map the long to a short text, so the dataset needs to be prepared as pairs of long text and its corresponding short summary. Automatic text summarization models based on deep learning architectures learn better on large datasets (Turc et al., 2019), so obtaining large and appropriate datasets is crucial for summarization tasks. Among published datasets, the most represented are in the English language. Furthermore, among English datasets, summarization research uses 55,65% of the publicly available dataset and 30,35% uses private datasets (Widyassari et.al., 2022).

The multilingual version of CNN/DailiyMail summarization dataset is a standardly used as the evaluation benchmark (Papers with Code, n.d.) available for five different languages (French, German, Spanish, Russian, Turkish). The Croatian translation of CNN/DailyMail is used here for the training of sequence-to-sequence abstractive summarization.

Existing Croatian neural models are trained on multilingual datasets. The CroSloEngualBERT model is trained in Croatian, Slovenian, and English languages (Ulčar & Šikonja, 2020) for NER, POS-tagging, and dependency parsing tasks. The model BERTić is trained in the Bosnian, Croatian, Montenegrin, and Serbian languages (Ljubešić & Lauc, 2021) for NER, POS-tagging, geo-location prediction, and commonsense causal reasoning tasks. The Context-aware Croatian Abusive Language (CoRAL) dataset (Shekhar et al., 2022) is used for detecting abusive language tasks. It contains user comments from one of the Croatian news portals (Shekhar et al., 2021). The Cro-CoV-cseBERT model automatically analyses the collected tweets in Croatian language, related to the communication about COVID-

19. Dataset is constructed of 10 000 manually annotated tweets (Babić et al., 2021).

The goal of this research is to obtain a dataset for training and testing the Croatian abstractive text summarization. For that purpose, a Croatian dataset is created using a machine-translated version of the well-known CNN/DailyMail English dataset. Then two different sequence-to-sequence models with attention mechanisms: long short-term memory (LSTM) and bidirectional long short-term memory (Bi-LSTM) are trained. The results of both models are evaluated with ROUGE 1, 2, L, and BERTSum evaluation metrics and compared with the same model trained in the original English language.

The paper is structured as follows: Section 2 overviews the sequence-to-sequence architecture and evaluation principles. The constructed dataset is elaborated on in Section 3, while Section 4 presents the results. The paper ends with a discussion.

2 Background architecture

2.1 Sequence-to-sequence

This work is based on a sequence-to-sequence with attention architecture (Sutskever et al., 2014). The architecture consists of encoder and decoder stacks of neural networks (Cho et al., 2014) that map input text to output text. Before the sequence is passed to the encoder, it is tokenized into words and converted into a word embedding. The output of the encoder is the context vector that goes to the decoder. The decoder predicts the next word based on information from the encoder and the previous predicted word. The attention mechanism additionally provides the decoder with information about relevant parts of the input sequence and the association between them (Fig.1).

Attention was originally applied to neural machine translation (NMT) tasks (Bahdanau et al., 2015), and then implemented in abstractive summarization (Rush et al., 2015; Nallapati et al., 2016). Attention enables focusing on the most relevant part of a source sentence where the information is concentrated, enabling the model to improve the prediction of the target word.

The neural networks in the encoder and decoder stack are layered on top of each other to provide better results (Suleiman & Awajan, 2020). They can utilize vanilla recurrent neural networks (RNN) (Nallapati et al., 2016), LSTM (Sutskever et al., 2014), gated recurrent unit (GRU) (Cho et al., 2014) or their bidirectional representatives (Kaichun et al., 2018; Prethee et al., 2022).

Analysis of the several approaches shows that recurrent neural networks with an attention mechanism and LSTM (Hochreiter & Schmidhuber, 1997; Cheng et al., 2016) are the most prevalent techniques for abstractive text summarization (Suleiman & Awajan, 2020).

While regular networks are trained in a forward direction, bidirectional networks (Schuster & Paliwal, 1997), are trained in both directions. Bi-LSTM (Graves et al., 2005) contains two LSTM layers: forward and backward LSTM, so they capture past and future information of the input word. The output of both layers is concatenated into functions such as average, sum, product or concatenation.

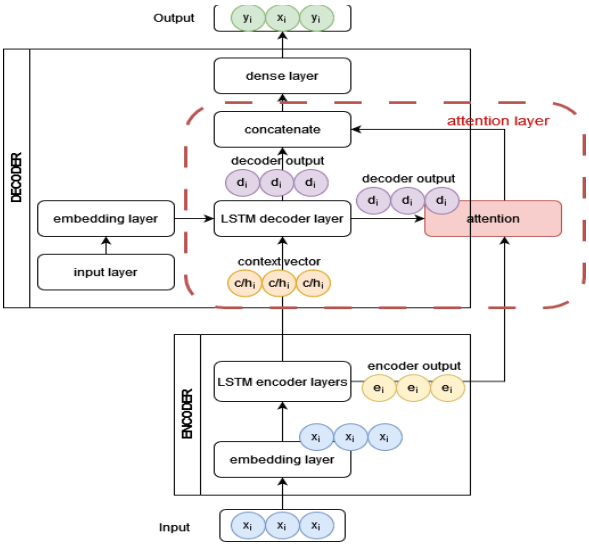

Figure 1. Encoder-decoder architecture with an attention mechanism (x_i=input vector, e_i=encoder hidden state, c/h_i= context vector, d_i=decoder output, y_i=output vector)

Transformers (Vaswani et al., 2017.) represent the other line of promising architectures for many text generation tasks. In this work, we limit the initial example to a sequence-to-sequence baseline, and we plan to train transformers in the following work.

2.2 Evaluation

The standard evaluation metric for summarization is Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin & Och, 2004). While ROUGE-N {N=1, 2, 3, …} calculates n-gram overlap between the machine-generated and reference summaries, ROUGE-L calculates the longest common subsequence (LCS) between the generated and reference summary so similar summaries have longer LCS. ROUGE and similar metrics correlate poorly with human evaluation (Novikova et al., 2017) because they do not consider the meaning and context of the evaluated text. BERTScore aims to overcome this drawback.

The BERTScore (Zhang et al., 2020) evaluation metric includes a contextual word embedding measure that calculates the similarity score between tokens in the original and the generated summaries. This metric is based on pre-trained BERT contextual embeddings (Devlin et al., 2019) and WordPiece tokenization of

input text (Wu et al., 2016). BERTScore uses multilingual WordPiece tokenization for non-English languages.

3 Dataset

Cable News Network (CNN)/DailyMail (Hermann et al., 2015; Nallapati et al., 2016) is an English-language dataset containing news articles, written by CNN and DailyMail journalists. The original version was created for the machine-reading and comprehension and question-answering tasks. One text in the dataset consists of id, article content, and highlights. *Id* is a string containing the hexadecimal formatted SHA1 hash of the URL where the story was retrieved from. *Article* is a string containing the text of the news article. *Highlights* is a string containing the summary of the article, written by the author of the article.

The dataset is divided into three separate parts: training, testing, and validation. The training subset contains 287,113; the test subset 11,490; and the validation subset 13,368 pairs of text and summary.

3.1 Croatian ML translated dataset

CNN/DailyMail dataset is commonly used for abstractive summarization tasks (Suleiman & Awaian, 2020). In this work, Google Machine Translation (GMT) is used to translate it into Croatian language (Google Cloud Translation, n.d.). The original division of the dataset into a training, testing and validation part is preserved and the resulting translations are prepared in JSON format, with UTF-8 encoding.

The analysis of the original and translated DailyMail/CNN dataset in English and Croatian languages is presented in Table 1.

Table 1. The average number of words in the text and summary in CNN/DailyMail in English and machine-translated Croatian datasets

CNN/DailyMail datasets	Avg. num. of words in the document	
	text	summary
English	659	49.9
Croatian MT	579.8	45.7

4 Experimental settings

In this work, we use a sequence-to-sequence architecture based on encoder-decoders with attention mechanisms. We trained two language models: one based on three layers of LSTM networks and another based on a Bi-LSTM network. Both models have an attention mechanism.

The LSTM and Bi-LSTM models were trained using the Croatian MT CNN/DailyMail dataset. Additionally, the LSTM model was also trained with the original English CNN/DailyMail dataset to compare the results

4.1 Preprocessing

Both Croatian MT and English datasets are initially preprocessed and tokenized. During preprocessing, non-alphabetic characters, multiple spaces, and single characters are removed from the training and validation text. Additional text cleaning is performed with the Python library SpaCy (Honnibal & Montani, 2017). There are available pipelines trained for specific languages. Pipeline *hr_core_news_lg* (Ljubešić et al., 2018) was trained on the Croatian language and *en_core_web_sm* (Weischedel et al., 2013; Choi & Palmer, 2012; Fellbaum, 2005) was trained on the English language.

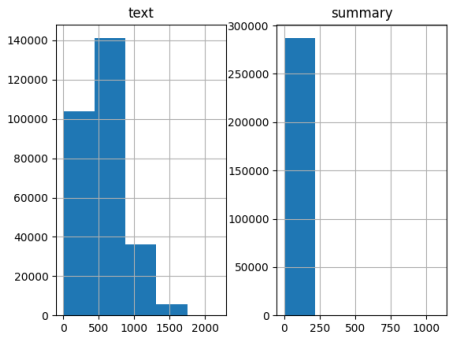

Figure 2. Distribution of words in DailyMail/CNN dataset. X-axis represents number of the words in the document. Y-axis represents number of the documents. Left: distribution in texts; Right: distribution in summary

During the preparation of the datasets for the training, we decided to set the maximum length of the texts to 800 words and summaries to 50 words. According to the word distribution (Fig.2) this seems to be an adequate threshold.

Table 2. Vocabulary size in texts and summaries

CNN/DailyMail datasets	Vocabulary size	
	text	summary
English	136,927	30,963
Croatian MT	305,163	64,445

Texts and summaries exceeding the maximum length are filtered out from the training dataset. Documents that have less than 800 words in text or 50 words in the summary are expanded by padding a zero up to the maximum length. After preparation, the English dataset contains 135,474 documents and the

Croatian dataset contains 171,107 documents that were used for training.

Two vocabularies are extracted per both datasets: one from the original texts and the other from the summaries. The vocabularies from the original text and the summary contain the most frequent words from the original text and the summary in both languages. The differences in the number of words in the vocabularies are displayed in Table 2. In contrast to the English, Croatian language has a rich morphology in which one word may have many forms (Seljan, 1999), and that explains the larger vocabulary of the Croatian dataset.

4.2 Language models

Two sequence-to-sequence models with attention mechanisms are compared. The encoder stack in the first LSTM model has three connected LSTM layers. The output of the last LSTM layer goes to the decoder side: the attention layer and the LSTM layer. The LSTM layer in the decoder gets a context vector from the encoder. The context vector is formed from the encoder hidden states and the attention weights. The outputs from the decoder's LSTM and attention layer go to the concatenation layer and from there to the last, dense layer. The dense layer has a softmax activation function that decides which token to generate next.

The encoder stack of the Bi-LSTM model has one Bi-LSTM layer. From this layer, outputs go to the decoder's concatenation layers and the attention layer. The concatenation layers merge inputs and pass them to the LSTM on the decoder side. The outputs from this layer and the attention layer are concatenated and passed to the dense layer to obtain the generated summary.

4.3 Training

Models were trained in the Tensorflow/Keras environment, using an AMD Ryzen Threadripper 3960X 24-Core Processor, CPU 2.2 GHz, and 256 GB of RAM. The hyperparameters for LSTM and Bi-LSTM models have the same values. The latent dimension was set to 300 and the embedding dimension to 200 tokens. The optimizer for models is RMSprop and the learning rate is set to 0.001. Details about the training are presented in Table 3.

For the Croatian dataset, the batch size is set to 250 and for the English dataset, the batch size is set to 400. A smaller batch size for the Croatian dataset was chosen because the model is trained with a large number of parameters and a larger batch size can exceed the memory limit.

Training is performed with maximum 50 epochs with an early stopping mechanism that is activated when the model does not improve results.

Table 3. Comparison of training data in different models and datasets (ES=early stopping)

Dataset and model	Num. of parameters	Batch size	Train. time	Num. of epochs
EN-LSTM	55,011,863	400	122h	33 (ES)
HR-LSTM	115,486,155	250	148h	26 (ES)
HR-Bi-LSTM	135,392,655	250	303h	50

5 Results

The models are tested during the inference phase with the testing subsets and evaluated using standard evaluation metrics ROUGE 1, 2, and L. Since the ROUGE metrics do not capture the semantics of the text, the BERTScore metric is used for additional evaluation. Table 4 shows the results of each model.

Table 4. Evaluation using ROUGE and BERTScore metrics

Dataset and model	Rouge 1 (F1)	Rouge 2 (F1)	Rouge L (F1)	BERTScore (F1)
EN-LSTM	20.29%	3.41%	15.10%	82.42%
HR-LSTM	16.91%	2.75%	12.58%	64.94%
HR - Bi-LSTM	18.71%	3.17%	13.22%	66.17%

The LSTM model trained on the Croatian MT dataset achieved the lowest results regardless of the used metrics. The Bi-LSTM model slightly improves the results over the LSTM baseline. Although the LSTM model on the English dataset shows better results compared to the models trained on the Croatian MT dataset, the overall results on ROUGE metrics are not impressive. In general, models achieve better results with BERTScore considering the content of texts. However, the BERTScore result on the Croatian MT dataset (64-66%) underperforms the English counterpart (82.42%). According to authors (Zhang et al., 2020), "for non-English language, the multilingual $BERT_{multi}$ is a suitable choice although BERTScore computed with this model has less stable performance on low-resource languages".

Despite modest summarization results, examples of the translated summary and the summary generated by LSTM and Bi-LSTM in the Croatian are given below:

Table 5. Examples of abstractive summarization with LSTM and Bi-LSTM models on the Croatian MT dataset

Translated summary	*pijani tinejdžer popeo se u nastambu s lavovima u zoološkom vrtu u zapadnoj indiji rahul kumar 17 potrčao je prema životinjama vičući ću ubiti lava srećom pao je u jarak prije nego što je stigao do lavova i spašen je*
Predicted summary LSTM	*slon je bio zoološkom vrtu za safari parku južnoj africi ali je pobjegao iz zoološkog vrta ali je preživio*
Predicted summary Bi-LSTM	*zoološkog vrta kini je životinje zoološkom vrtu central parku kina je upozorila da je životinja na putu za dobrotvorne svrhe ali je umro nakon što je pobjegao iz zoološkog vrta*
Translated summary	*reanne evans suočila se s kenom dohertyjem u kvalifikacijama za svjetsko prvenstvo doherty je osvojio svjetsko prvenstvo 1997 evans je izgubio prvi okvir 71-15 protiv dohertyja ali dudley native se izborio i poveo 4-3 ken doherty, međutim uspio je zatvoriti zadivljujuće natjecanje 10-8*
Predicted summary LSTM	*trener je rekao da je bio na putu za trening nakon što je izgubio 3 0 od njemačke finalu svjetskog prvenstva 2010 ali je uspio pobijediti na turniru na svjetskom prvenstvu 2014 ali je uspio pobijediti na turniru na igrama londonu*
Predicted summary Bi-LSTM	*evans je pobijedio 6 5 7 5 6 5 7 5 polufinalu na australian openu evans je počeo igrati za pobjedu od 7 6 7 6 7 6 7 6 7 6 7 6 1994 sada je pobijedila 6 0 7 6 7 5 polufinalu*

6 Discussion and Conclusion

The summarization task relies on a large amount of data that consists of pairs of text and summary. It represents a laborious challenge to obtain an adequate dataset for additional language such as Croatian. To acquire the dataset, we used Google to machine translate CNN/DailyMail into the Croatian language.

On Croatian MT CNN/DailyMail dataset, the baseline sequence-to-sequence models - LSTM and BiLSTM with attention mechanism are trained. The results are evaluated on ROUGE 1, 2, L, and BERTScore metrics and then compared with the evaluation performed on the original English dataset.

ROUGE is a simple metric that compares exact words between reference and generated summaries and does not give information about the semantic similarity of the words. For example, ROUGE cannot recognize that different words can represent the same term. ROUGE metrics show that sequence-to-sequence models in general do not infer good results, yet the English model achieves slightly better results than the Croatian. That is expected since the number of parameters is higher due to the larger vocabulary and for the training of Croatian summarization the larger quantities of texts are needed.

The results on the English dataset are much better than the results on the Croatian MT dataset. Regardless of the results, generated summaries are not of the desired content, but topics are well captured. The model correctly infers the topic but misses generating the correct information. Still, this initial attempt is indicative and motivational, that with a lot of improvements both in the selected architecture and in the dataset, we can progress towards the adequate solution for the abstractive text summarization in the Croatian language.

To conclude, the translated version of CNN/DailyMail can serve as a shortcut for training the initial summarization model for Croatian.

As the next step, the Croatian MT dataset will be extended with additional data and used to train transformer models. Transformers made huge progress in the field of abstractive summarization and regularly reached much higher evaluation scores than sequence-2-sequence models, however, they require a lot of computing resources and a huge amount of data to achieve better results.

Acknowledgment

This work has been fully supported by the University of Rijeka under project number uniri-drustv-18-20.

References

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K., Text Summarization Techniques: A Brief Survey, *International Journal of Advanced Computer Science and Applications (ijacsa)*, 8(10), 2017. http://dx.doi.org/10.14569/IJACSA.2017.081052

Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Matešić, M., Meštrović, A. (2021). Characterisation of COVID-19-Related Tweets in the Croatian Language: Framework Based on the Cro-CoV-cseBERT Model. In *Applied Sciences, 2021*. 11(21), 10442; https://doi.org/10.3390/app112110442

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align

and translate. In *3rd International Conference on Learning Representations (ICLR'15),* Bengio Yoshua and LeCun Yann (Eds.).

Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. In *2016 Conference on Empirical Methods in Natural Language Processing*, pp. 551–561, Las Vegas, NV, USA.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Empiricial Methods in Natural Language Processing (EMNLP 2014)*, October. arXiv preprint arXiv:1406.1078.

Choi, J.D., &Palmer, M. (2012). Guidelines for the CLEAR Style Constituent to Dependency Conversion, *Technical report 01-12: Institute of Cognitive Science*, University of Colorado Boulder

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, pp 4171–4186. arXiv preprint arXiv:1810.04805.

Fellbaum, C. (2005). WordNet and wordnets. In *Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition.* pp 665-670. Oxford: Elsevier

Google Cloud Translation. (n.d.). https://cloud.google.com/translate

Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification and recognition. In *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005, Part II 15* pp 799-804. Springer Berlin Heidelberg.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9(8),* pp 1735-1780, November 1997, doi:-10.1162/neco.1997.9.8.1735.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1), pp 411-420.

Lin, C. Y., & Och, F. J. (2004). Looking for a few good metrics: ROUGE and its evaluation. In *4th Ntcir Workshops*. 2004, pp. 1-8.

Ljubešić, N., Agić, Ž., Klubička, F., Batanović, V., & Erjavec, T. (2018). Training corpus hr500k 1.0., *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042, http://hdl.handle.net/11356/1183.

Ljubešić, N., & Lauc, D. (2021). BERTić--The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In *8th Workshop on Balto-Slavic Natural Language Processing*, pp 37–42, Kiyv, Ukraine. arXiv preprint arXiv:2104.09243.

Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL-16 - 20th SIGNLL Conference on Computational Natural Language Learning*, pp 280–290. https://doi.org/10.18653/v1/k16-1028

Novikova, J., Dušek, O., Curry, A. C., & Rieser, V. (2017). Why we need new evaluation metrics for NLG. In *2017 Conference on Empirical Methods in Natural Language Processing*. arXiv preprint arXiv:1707.06875.

Papers with Code. (n.d.). Document Summarization. https://paperswithcode.com/task/document-summarization

Rush, M., Chopra, S., & Weston, J., (2015). A neural attention model for abstractive sentence summarization. In *2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, Boston, MA, USA.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11), pp 2673-2681

Seljan, S. (1999). Lexical-Functional Grammar of the Croatian Language. In *22nd International Convention MIPRO '99* - CIS Computers in Intelligent Systems (CIS) Rijeka: Liniavera.

Shekhar, R., Karan, M., & Purver, M. (2022). CoRAL: a Context-aware Croatian Abusive Language Dataset. In *Findings of the ACL: AACL-IJCNLP*. arXiv preprint arXiv:2211.06053.

Shekhar, R., Pranjic, M., Pollak, S., Pelicon, A., & Purver, M. (2021). 24sata news comment dataset 1.0., *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042, http://hdl.handle.net/11356/1399.

Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. In *Mathematical problems in engineering, 2020*, pp 1-29. https://doi.org/10.1155/2020/9365340

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems, 27*. pp 3104-3112. arXiv preprint arXiv:1409.3215

Turc, I., Chang, M. W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962.

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020*, Proceedings 23. pp 104-111. Springer International Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, pp 5998–6008.

Weischedel, R., Palmer, M., Marcus, M., Hovy,E., Pradhan, S., Ramshaw, L., …Houston, A. (2013). OntoNotes Release 5.0. *Philadelphia: Linguistic Data Consortium.* https://doi.org/10.35111/xmhb-2b84

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., & Affandy, A. (2022). Review of automatic text summarization techniques & methods. In *Journal of King Saud University-Computer and Information Sciences*, 34(4), pp 1029-1046.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., …Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*. arXiv preprint arXiv:1904.09675.