Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Summarization of Slovak News Articles

Bachelor Thesis

2023

Viktória Ondrejová

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Summarization of Slovak News Articles

Bachelor Thesis

Study Programme:	Data Science
Field of Study:	Computer Science and Mathematics
Department:	Department of Computer Science
Supervisor:	Mgr. Marek Šuppa

Bratislava, 2023
Viktória Ondrejová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Viktória Ondrejová
Študijný program:	dátová veda (Medziodborové štúdium, bakalársky I. st., denná forma)
Študijné odbory:	informatika matematika
Typ záverečnej práce:	bakalárska
Jazyk záverečnej práce:	anglický
Sekundárny jazyk:	slovenský

Názov: Summarization of Slovak News Articles
Sumarizácia slovenských novinových článkov

Anotácia: Sumarizácia textu je jedným zo základných problémov spracovania prirodzeného jazyka, a to tak v jeho extraktívnej, ako aj abstraktnej forme. Hoci sa na tento problém zameriava značná časť publikovaných prác, pričom významné zlepšenia sa uvádzajú najmä v prípade modelov založených na architektúrach hlbokých neurónových sietí, tieto sa zvyčajne vykonávajú v jazykoch s vysokými zdrojmi, ako je angličtina. Existujú aj iniciatívy, ktoré sa zaoberajú sumarizáciou textov viacjazyčným spôsobom, ale pokiaľ je nám známe, modely sumarizovania textov neboli testované v kontexte slovenského jazyka. Dá sa to vo všeobecnosti vysvetliť nedostatkom dostupných slovenských korpusov vhodných na textovú sumarizáciu.

Cieľ: Ciele bakalárskej práce zahŕňajú (ale nie sú obmedzené na)
- analýzu aktuálne dostupného slovenského sumarizačného súboru(ov)
- vytváranie a/alebo zber nových súhrnných dátových súborov v slovenskom jazyku
- tréning základných modelov na pripravených súboroch údajov
- vyhodnotenie najmodernejších viacjazyčných sumarizačných modelov na pripravených datasetoch

Literatúra: - EL-KASSAS, Wafaa S., et al. Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 2021, 165: 113679. (https://doi.org/10.1016/j.eswa.2020.113679)
- HASAN, Tahmid, et al. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. arXiv preprint arXiv:2106.13822, 2021. (https://arxiv.org/pdf/2106.13822.pdf)
- SUPPA, Marek; ADAMEC, Jergus. A summarization dataset of Slovak news articles. In: Proceedings of the 12th language resources and evaluation conference. 2020. p. 6725-6730. (https://aclanthology.org/2020.lrec-1.830/)

Vedúci:	Mgr. Marek Šuppa
Katedra:	FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry:	prof. Ing. Igor Farkaš, Dr.

Spôsob sprístupnenia elektronickej verzie práce:
prípustná pre vlastnú VŠ

Comenius University Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Viktória Ondrejová
Study programme:	Data Science (Joint degree study, bachelor I. deg., full time form)
Field of Study:	Computer Science Mathematics
Type of Thesis:	Bachelor´s thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Summarization of Slovak News Articles

Annotation: Text summarization is one of the fundamental problems in natural language processing, both in its extractive as well as abstractive form. Although a significant body of previously published works focus on this problem, with significant improvements being reported especially for models based on Deep Learning architectures, these are usually done on high-resource languages such as English. There are also initiatives which deal with text summarization in a multilingual manner but to the best of our knowledge, the text summarization models haven't been tested in the context of Slovak language. This can be generally explained by the lack of available Slovak corpora suitable for text summarization.

Aim: The goals of the bachelor's thesis include (but are not limited to)
- analysis of the currently available Slovak summarization dataset(s)
- creation and/or collection of new summarization datasets in Slovak language
- training baseline models on the prepared datasets
- evaluation of state-of-the-art multilingual summarization models on the prepared datasets

Literature:
- EL-KASSAS, Wafaa S., et al. Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 2021, 165: 113679. (https://doi.org/10.1016/j.eswa.2020.113679)
- HASAN, Tahmid, et al. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. arXiv preprint arXiv:2106.13822, 2021. (https://arxiv.org/pdf/2106.13822.pdf)
- SUPPA, Marek; ADAMEC, Jergus. A summarization dataset of Slovak news articles. In: Proceedings of the 12th language resources and evaluation conference. 2020. p. 6725-6730. (https://aclanthology.org/2020.lrec-1.830/)

Supervisor:	Mgr. Marek Šuppa
Department:	FMFI.KAI - Department of Applied Informatics
Head of department:	prof. Ing. Igor Farkaš, Dr.

Electronic version available:
prípustná pre vlastnú VŠ

Acknowledgments:

Abstrakt

Kľúčové slová:

Abstract

Keywords:

Contents

Introduction

Chapter 1

Summarization

In this section, we will focus on summarization itself, its history, classification, and related work for non-English languages.

1.1 Early Work

Text summarization as a task of Natural Language Processing (NLP) has been around for quite some time. Its beggings date back to the 50s–60s when NLP as a whole gained popularity during World War II. Iin the first works researchers studied the following techniques:

- Position in the text: sentences in the privileged locations
- Lexical cues: adjectives (important, hardly, impossible) near the significant words
- Location of sentences: first and last sentences usually contain topic information

Even though these approaches are relevant, they extremely rely on the format and style of the summarized text. Techniques used in this period were very much limited by their time since neither large corpora of texts, effective NLP models, nor powerful computers existed. With the invention of the Web and new easy-to-implement methods, research in the text summarization problem began to rise in popularity. With the advent of the new millennium and neural networks, important development in the NLP area was done.

1.2 Classification

The main motivation for dealing with text summarization problem is the amount of information we are overwhelmed with on the daily basis. Now, more than ever, it's so easy to write a news article but many of them are unnecessarily long, repeat the same idea, or contain redundant information that wastes the reader's time. As everything done manually, manual summarization made by humans is costly and time-consuming when compared to Automatic Text Summarization (ATS). Each researcher can interpret summarization slightly differently, but generally, we can agree that while summarizing, we want to extract the most important parts of the original text and produce the shortest possible version of it while preserving context and providing readability, comprehensibility, and integrity of the original text. Summarization can by classified into two main approaches: Extractive and Abstractive.

1.2.1 Extractive Summarization

In short, when we use the extractive summarization approach, we select and extract the most important words from the input text. Firstly, we need to pre-process the input text. Then we can process it by creating a representation of text, scoring and ranking sentences using metrics, and extracting high-scored sentences. Finally, we need to post-process our extracted sentences.[3]

Because of its nature, the extractive summaries are far from perfect. They are much more prone to redundancy, produce longer summaries, lack cohesion and semantics because of the incorrect links between sentences, and much more. Despite all of that, summarization by extraction is still a good approach because of its simplicity and higher accuracy because the words in the final summary already exist in the input.

Even though, the first summarization models were based on the extractive approach, with time it got sophisticated and it's still used to this day. We will focus on the current state-of-the-art models.

Transformers

Many state-of-the-art models are based on the Transformer architecture, first introduced in 2017. [10] The Transformer aims to solve sequence-to-sequence tasks, like Recurrent Neural Networks (RNN), which were unable to deal with long-range dependencies and unable to perform parallelization. The model was built to use self-attention while computing the input and output representation. The Transformer model consists of the encoder and decoder parts. The encoder receives each component of the input sequence and encodes it into a vector carrying the context information about the whole sequence. This context vector is sent to the decoder, which reads the vector, understands the context, and creates a meaningful output. The most important part of the whole architecture, the context vector, contains all the information of the original sequence. To prevent a bottleneck and achieve better encoding, the self-attention mechanism is introduced. In self-attention, the encoder not only passes all of the hidden states to the decoder [1] but also looks for context clues in other elements of the sequence as it processes them by calculating a weighted combination of all other elements (including elements that appear later in the sentence). These calculations are done multiple times in parallel. That way the self-attention tracks associations from the input sequence, this can be used to extract an understanding of each of the processed elements.

In the original paper, the transformer consisted of 6 encoders stacks and 6 decoders stacks. In a typical encoder-decoder architecture each encoder consists of a self-attention layer and a feed-forward neural layer. Similarly, the decoder is composed of three parts: a self-attention layer, a decoder attention layer, and a feed-forward neural net. When processing a sequence, the first encoder takes the input vectors, processes them with the self-attention layer, and passes them to the feed-forward neural net. The result is passed to the next encoder and so on until the final encoder sends the information to all the decoder attention layers where a similar process is done.

-bidirectional

BERT-based Models

As its name implies, Bidirectional Encoder Representations from Transformers (BERT) use a Transformer mechanism to complete multiple NLP tasks. Since BERT's goal is to generate a language model, it uses only the encoder mechanism. It was trained using Transfer learning, where the model uses the knowledge gained in the pre-training and applies it to a different but related problem. BERT was trained on Mask Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks.

In the MLM training, BERT received word sequences as input and each of the sequences had 15% of the words replaced with a [MASK] token. To predict the output words, BERT first needs to add a classification layer on top of the encoder output. Then multiply the output vectors by the embedding matrix and therefore transforms them into the vocabulary dimension finally, it calculates the probability of each word in the vocabulary by softmax. Because of the BERT loss function, which takes into consideration only the prediction of the masked values, the model converges slower than directional models.[4]

In the NSP training, BERT received pairs of sequences as input and learned to predict if the two sequences are from the same document. During training, 50% of the inputs are from the same documents, while in the other 50%, the second sentence is randomly chosen from the corpus. To distinguish the two sentences in training, each of the sequence is processed by inserting [CLS] and [SEP] tokens at the beginning of the first sentence and the end of each sentence respectively. Then, a sentence embedding is added to each token as an indicator of Sentence A or Sentence B. Finally, information about the position in the sequence - positional embedding is added to each token. Input

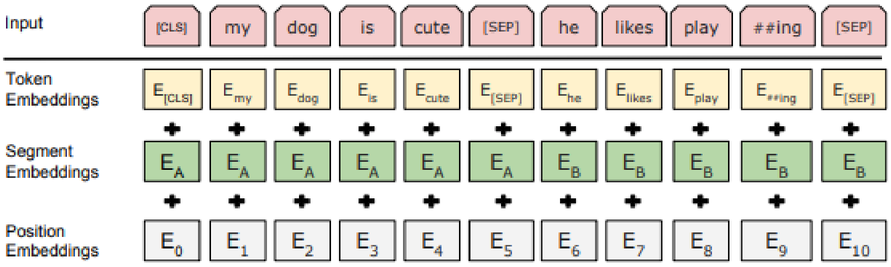

Figure 1.1: Representation of the BERT input [2]

processed this way then enters the model, where the entire input sequence firstly goes through the Transformer model, then, using a simple classification layer, the output of the [CLS] token is transformed into a 2x1 shaped vector. And finally, the probability of the second sentence being from the same document is calculated using softmax.[4]

Both trainings were done simultaneously to minimize the combined loss function of the two strategies as shown in Figure 1.1. Thanks to the transfer learning, which BERT takes advantage of, a model trained on this tasks can be fine-tuned for nearly any NLP tasks with minimal changes.

???The multilingual version of BERT is MBERT, which was trained and used with 104 languages. -sampling from different languages

1.2.2 Abstractive Summarization

The summary generated by the abstractive approach consists of paraphrased main ideas of the input text using different vocabulary than in the original text. It also pre-processes, processes, and post-processes the input text but instead of scoring and extracting sentences, it generates a summary using natural language generation techniques.

This results in better summaries, that are closer to those generated by humans. It can paraphrase and compress. Therefore, producing an even shorter summary than with the extractive approach. But generating a good abstractive summary is very difficult to achieve because the generator itself requires a full interpretation of the input text and it's still in research. Models are restricted by used representations and can't fully work with out-of-vocabulary words. [3] Even though research in the abstractive approach advanced significantly, it's still difficult to generate a good abstractive summary.

T5-based Models

The Text-to-Text-Transfer-Transformer model (T5) attempts to combine all NLP tasks into the "Text-to-Text" format. The Text-to-Text framework, used for this purpose, uses the same model, loss function, and hyperparameters on all NLP tasks.[9] Therefore, inputs need to be modeled in such a way that the model will recognize the specific task and produces text output. To make this possible, a task-specific text prefix is added at the beginning of the original input. For example, for summarizing, the "summarize:" prefix is added to the input sequence. To achieve the shortest sequences possible, T5 uses a denoising objective where 15% of the original sequence, including both individual words and sequence of words, are masked for the model to predict them.[9] The T5 model consists of the standard transformer model: 12 pair blocks of encoder-decoder and its size and configuration are similar to the "BERTBASE" stack and was trained on a C4 dataset. [7] The T5's multilingual variant is mT5, pretrained on an mC4 dataset covering 101 languages, Slovak including. [11]

BART-based Models

description of mBART

1.3 Related Work

Because the majority of the research is focused on the English language, non-English languages are at a significant disadvantage. Many models and methods used for summarization are designed specifically for English because of its universality. Nonetheless, over the last two decades, multilanguage text summarization was the object of study for several researchers. One of the earliest works was MEAD which could summarize clusters of news articles for both English and Chinese using an extractive approach.[6] One of the many ways how to summarize low-resource languages is to translate datasets to English and summarize and translate it back to the original language. This summarization is significantly underperforming because of the translation biases that add to the difficulty of summarization.

Latest extractive state-of-art methods such as BERT now exist for various languages and its multilingual version M-BERT is often extended to low-resource ones. More recently, with the creation of new multilingual datasets, which are required for training models, the abstractive approach began to rise in popularity. One of the biggest datasets for abstract summarization is XL-Sum which consists of 44 languages ranging from low to high-resource, Slovak language including. This encourages various languages to start their research in NLP.

Slovak summarization as an NLP task was neglected for a long period of time. Only recently, progress and research in this area have begun to emerge. The first article written on Slovak summarization was A Summarization Dataset of Slovak News Articles where Šuppa and Adamec created the SME dataset, the first Slovak dataset for summarization which consists of more than 80 000 news articles from Slovak newspaper SME. [8] In their work, the SME dataset was used in extractive summarization using the M-BERT model and evaluated with modified ROUGE metrics better suitable for the Slovak language.

Another significant milestone for Slovak NLP was the creation of the SlovakBERT developed by the Kempelen Institute of Intelligent Technologies (KInIT) and Gerulata Technologies. Slovak-BERT is a large-scale transformers-based Slovak masked language model using more than 19GB of Web-crawled Slovak text. [5] Although multilingual models exist, the Slovak-only language model can lead to better results and more compute and memory-wise efficient processing of the Slovak language.

In this thesis, we aim to create an even larger dataset containing news articles from multiple newspapers and test both extractive and abstractive methods.

Chapter 2

Data

This section is all about the data used in this thesis. We will describe existing datasets, the creation and cleaning of the new dataset.

2.1 Existing datasets

Sme dataset and some multilanguage datasets.

2.2 New dataset

Something like: As we "saw", current state of Slovak summarization calls for new datasets to be created. And so on.

2.2.1 Data sources

Here we will describe sources, from which we obtained new data.

2.2.2 Extraction

How we scraped the data.

2.2.3 Data cleaning

What needed to be deleted/modified.

2.2.4 Cool name for new dataset

Description and stats of the created dataset. Structure of the dataset.

Chapter 3

Experiments and evaluation

In this section we will focus on the experiments themselves, we describe the mohel fine-tuning and various problems that come with it. Next, we will evaluate our experiments with metrics that had to be adjusted for the Slovak language.

3.1 Experiments

This will be divided into more subsections.

3.2 Evaluation

Results

Here we will go over or results and describe them.

Conclusion

Bibliography

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[3] Wafaa El-Kassas, Cherif Salama, Ahmed Rafea, and Hoda Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679, 07 2020.

[4] Rani Horev. Bert explained: State of the art language model for nlp, 2018.

[5] Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. Slovakbert: Slovak masked language model, 2021.

[6] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. 2000.

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[8] Marek Suppa and Jergus Adamec. A summarization dataset of Slovak news articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6725–6730, Marseille, France, May 2020. European Language Resources Association.

[9] T5: Text-To-Text Transfer Transformer. T5: Text-to-text transfer transformer, 2020.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[11] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.