SlovakSum: A Large Scale Slovak Summarization Dataset

Viktória Ondrejová, Marek Šuppa
Department of Applied Informatics, FMFI UK   Cisco Systems
Mlynska dolina, 842 48   Bratislava
ondrejova18@uniba.sk, marek.suppa@fmph.uniba.sk

Abstract

The ability to automatically summarize news articles has become increasingly important due to the vast amount of information available online. Together with the rise of chatbots , Natural Language Processing (NLP) has recently experienced a tremendous amount of development. Despite these advancements, the majority of research is focused on established well-resourced languages, such as English. To contribute to development of the low resource Slovak language, we introduce SlovakSum, a Slovak news summarization dataset consisting of over 200 thousand news articles with titles and short abstracts obtained from multiple Slovak newspapers. The abstractive approach, including MBART and mT5 models, was used to evaluate various baselines. The code for the reproduction of our dataset and experiments can be found at https://github.com/NaiveNeuron/slovaksum.

Keywords: SlovakSum, summarization dataset, document summarization, Slovak

1. Introduction

Automatic text summarization is an important task in the field of NLP, that aims to create a shorter version of the input text while preserving the main information, context, and integrity. With the rapid growth of the amount of online text, the summarization task has become an essential tool for various applications such as information retrieval or headline generation. As many state-of-the-art models are based on neural networks, large corpora is needed to train and evaluate these models effectively. Many significant datasets with hundreds and thousands of documents exist, but their content is mainly in the English language. With regards to non-English languages, the need for a large corpora is even bigger, as the available data is generally small or non-existent.

The aim of this paper is to present a large Slovak news summarization dataset, consisting of articles from multiple popular Slovak news websites. The dataset consists of headlines, short abstracts, and full text. To evaluate the performance of two abstractive summarization model baselines, we will use the standard ROUGE metric Lin and Hovy (2003).

2. Related Work

2.1. Summarization Methods

There are two main approaches for automatic summarization: *extractive* and *abstractive*. An extractive summary consists of the most important words or sentences from the input text chosen by a specific metric. This was used in the very first automatic summarization paper Luhn (1958), where the authors attempted to create an abstract of technical papers and news articles by selecting sentences with the highest significance. For many years, this was the main approach studied by researchers in the text summarization community. The focus is nowadays shifting towards the abstractive approach, where the generated summary is a paraphrase of the main points of the document El-Kassas et al., 2021.

The current state-of-the-art summarization models are based on the Transformer architecture introduced in Vaswani et al. (2017), which significantly advanced the NLP field in general. This architecture is used for both summarization approaches. The best performing extractive models are BERT-based, including `BERTSUM` Liu (2019) which employs a single encoder and a classification layer for ranking sentences and `HIBERT`, introduced in Zhang et al. (2019), an extension of the `BERTSUM` with a hierarchical attention mechanism to capture the relationships between document-level and sentence-level representations. In the abstractive approach, the Transformer architecture is used as a building block of `BART` Lewis et al. (2019), `T5` Raffel et al. (2020) and `PEGASUS` Zhang et al. (2020) that also incorporates multi-head attention and residual connections when producing a summary.

2.2. Datasets

To train and evaluate summarization models, various datasets have been created over the years, mainly in the context of English and other high-resource languages. In the past, the dataset from Document Understanding Conference (DUC) Over et al. (2007) was the most widely used corpora for single document summarization. It consists of 500 news articles and its biggest advantage is that it contains 4 different human-written

LREC-COLING 2024, pages 14916–14922
20-25 May, 2024.

Website	Documents	
	Number	Percentage
dennikn.sk	33,014	15.83
aktuality.sk	118,753	56.95
hnonline.sk	35,770	17.15
pluska.sk	20,982	10.07
Total	208,519	

Table 1: Number of articles from individual newspapers.

reference summaries for each article. Due to their limited size, however, it has become unfeasible to train neural network-based models on them, leading to a demand for substantially larger corpora. This resulted in the creation of datasets, that are to this day used for testing English summarization approaches: CNN/DailyMail Hermann et al. (2015), NY Times Durrett et al. (2016), and XSum Narayan et al. (2018). These large-scale datasets contain hundreds of thousands of news articles with reference summaries.

Besides English text summarization datasets, the most notable corpora are the XL-Sum Hasan et al. (2021), MLSUM Scialom et al. (2020) and the MultiLing datasets. For instance, the MultiLing 2017 dataset Giannakopoulos et al. (2017) covers 41 languages, including Slovak, albeit on a smaller scale. The closest works comparable to ours is Straka et al. (2018), in which the authors introduced a large-scale news summarization dataset for Czech called SumeCzech. To the best of our knowledge, the only somewhat comparable work for Slovak would be the SMESum dataset introduced in Suppa and Adamec (2020). In contrast with our work, however, this dataset is smaller in scale, contains only documents that originate from a single source (the SME.sk news portal) and only evaluates extractive baselines.

2.3. Metrics

Since the beginning of Automatic Summarization as an NLP task, many metrics have been proposed for its evaluation. The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric Lin and Hovy (2003) remains the most commonly used evaluation metric for summarization systems. Initially proposed as an English-specific recall-based metric, it employs English stemmers, stop words and synonyms. However, this English-specificity constitutes a significant limitation for non-English summarization tasks. The authors of Straka et al. (2018) suggest a language-agnostic approach called $ROUGE_{RAW}$ as a solution to this problem. This alternative approach does not rely on stemmers and does not consider stop words or synonyms.

Website	Distribution		
	Train	Validation	Test
aktuality.sk	57.58	57.74	57.61
dennikn.sk	15.86	16.11	15.96
hnonline.sk	16.86	16.75	16.97
pluska.sk	9.58	9.44	9.44

Table 2: The percentage proportion of the training, validation and testing splits across the websites.

3. The Dataset

The main motivation behind our work is to create large Slovak summarization corpora that would be sufficient for the application of deep learning methods. This implies that the final dataset must already contains a reference summary within itself. We chose the Slovak online news portals `dennikn.sk`, `aktuality.sk`, `hnonline.sk` and `pluska.sk` based on their popularity[1] in order to cover a wide range of Slovak newswire text.

3.1. Extraction and Preparation

To extract the articles from the aforementioned online newspapers, we initially acquired all their known Wayback Machine URLs using `gau`[2]. Before extraction, duplicate links were removed and the URLs were filtered with keywords targeting images, API requests, comments, etc. Subsequently, the Scrapy framework [3] and the Wayback Availability JSON API[4] were combined to extract the headline, abstract and text from the specific HTML tags. Each scraper had to be adapted for a particular year because the HTML structure changed over time. In the final pre-processing step we deleted text resembling advertisement, paywall, HTML tags, and other non-textual content, such as chunks of JavaScript.

3.1.1. Structure

The final dataset consists of a headline, a short abstract, text, and the unique Wayback URL. As Table 1 shows, the majority of the dataset is comprised of `aktuality.sk`'s articles, followed by `hnonline.sk`, `dennikn.sk` and finally `pluska.sk`. The dataset was split into train/valid/test sets in the 8:1:1 ratio. The distributions of websites for all three sets can be found in Table 2. Table 3 shows a comparison of our dataset with the `SMESum` and other standard English datasets. The presented numbers suggest that the SlovakSum dataset's size is approaching that of large scale datasets

[1] The selected news portals were among the top 5 most popular at the time of dataset extraction, as per https://rating.gemius.com/sk/tree/112

[2] https://github.com/lc/gau

[3] https://scrapy.org/

[4] https://archive.org/help/wayback_api.php

Dataset	no. of documents			mean document length		mean summary length		vocabulary size	
	train	valid	test	words	sentences	words	sentences	document	summary
CNN	90,266	1,220	1,093	760.50	33.98	45.70	3.59	343,516	89,051
DailyMail	196,961	12,148	10,397	653.33	29.33	54.65	3.86	563,663	179,966
NY Times	589,284	32,736	32,739	800.04	35.55	45.54	2.44	1,399,358	294,011
XSum	204,045	11,332	11,334	431.07	19.77	23.26	1.00	299,147	81,092
SMESum	64,001	8,001	8,001	339.09	18.08	23.61	2.16	423,877	110,720
SlovakSum	166,815	20,852	20,852	238.44	17.21	20.03	2.58	864,185	230,289

Table 3: Comparison of the **SlovakSum** dataset with the **SMESum** dataset and some of the most popular English summarization datasets. The first column shows the number of documents in the train/valid/test sets. In the following two columns, the mean lengths (measured in both words and sentences) of the document and summary are displayed. Finally, in the last column, are the vocabulary sizes for each dataset. Values for the English datasets have been obtained from Narayan et al. (2018). Similarly, the **SME**'s values have been taken from its original paper Suppa and Adamec (2020).

in English, such as CNN/DailyMail or XSum. We also note that the vocabulary of the dataset is larger that of SMESum, which suggests that the dataset might be able to capture more variety than prior work. When compared to other datasets, the mean document length and the mean summary length of SlovakSum are smaller than in any of the other datasets, in some cases by as much as 70%. We attribute this fact to the presence of `aktuality.sk`'s articles, which are generally shorter, and comprise more than a half of the dataset (see Table 1).

The quantitative statistics of the headline, abstract, and text are displayed in Table 4. Compared to the `SumeCzech`'s statistics, the documents are quite short and are more similar to the `SMESum` numbers.

	Q1	Median	Q3	Mean	SD
`SumeCzech`					
document	265	378	553	470.1	365.3
summary	33	42	51	42.2	14.8
`SMESum`					
document	175	259	402	339.09	323.54
summary	19	22	26	23.61	6.25
`SlovakSum`					
document	127	185	293	238.44	169.06
summary	12	18	26	20.03	9.91

Table 4: Quantitative statistics of lengths of headlines and abstracts in words. Q1 and Q3 denote the first and the third quartile, respectively. Additionally, the standard deviation is shown in the SD column.

4. Experiments and Evaluation

4.1. Baselines

To put the results of our experiments in perspective, we firstly established and evaluated a selection of simple baselines.

One of the most popular baseline in automatic summarization is the `LEAD` method Nenkova (2005). The `LEAD` selects the first 3 sentences from the beginning of an original document to represent the output summary. With these sentences serving as a prediction, the $ROUGE$ F1 scores are computed.

Similarly, the `RANDOM` methods selects three random sentences as the generated summary, on which are then the $ROUGE$ F1 scores computed.

4.2. BART-based Model

In order to establish a strong baseline for our presented dataset, we used the BART language model (Bidirectional and Auto-Regressive Transformer) introduced in Lewis et al. (2019). BART is based on the Transformer architecture using a combination of bidirectional and auto-regressive methods. It was trained on large amounts of unannotated text using a denoising autoencoder architecture, where the model was trained to reconstruct the original input from a corrupted version.

To address the issue of the Slovak input data, we used the Multilingual BART (`mBART`) Liu et al. (2020), specifically its extended mBART-50[5] version Tang et al. (2020) created by multilingual finetuning. Despite being trained on 50 low-resource languages, the Slovak language was not included in the training. However, owing to the morphological similarities with the Czech language and many more languages present in its training corpora, we decided to evaluate our dataset on this model.

The `mBART` is based on a standard sequence-to-sequence Transformer architecture Vaswani et al. (2017), with 12 encoder layers, 12 decoder layers, 1024 hidden units, and approximately 670 million parameters. We implemented this model using HuggingFace's PyTorch Transformers Wolf et al. (2020) with all of the input text being tokenized by the `MBart50Tokenizer`, specifically designed for the `mBART` model. The model was fine-tuned for 90,000 steps on a single GPU with both batch size and number of epochs set to 8.

[5] Also known as `facebook/mbart-large-50`

	$ROUGE$			$ROUGE_{RAW}$		
Model	$R-1$	$R-2$	$R-L$	$R-1$	$R-2$	$R-L$
LEAD	18.48	6.64	13.44	14.6	4.2	11.0
RANDOM	16.22	5.19	11.71	13.0	3.4	9.9
mBART$_8$	21.73	10.58	18.37	20.8	9.0	18.5
mT5$_{11}$	16.52	7.24	14.59	14.8	6.0	13.5

Table 5: The $ROUGE$ and $ROUGE_{RAW}$ F1 scores of used abstractive models evaluated after various number of epochs (the number in subscript).

4.3. T5-based Model

The second model used for evaluation was the T5: Text-to-Text Transfer Transformer model introduced in Raffel et al. (2020). Taking advantage of transfer learning, this large-scale model was trained simultaneously on a total of 18 NLP tasks. For each task, the model takes textual input and generates a corresponding output text. Specifically, we used the multilingual T5 version introduced in Xue et al. (2021). While being limited with our GPU size, we used the mT5-BASE pre-trained on the mC4 dataset Xue et al. (2021) covering 101 languages, including Slovak. The mT5 uses the same standard Transformer architecture with 580 million parameters. Just like mBART, the mT5 is also easily implemented by the HuggingFace's PyTorch Transformers library Wolf et al. (2020), and was fine-tuned for 10 epochs.

4.4. Evaluation

To evaluate model performance on our dataset, we used the ROUGE metric presented in Lin and Hovy (2003). The ROUGE-1 (unigrams) and ROUGE-2 (bigrams) represent the n-grams overlap between the generated and the reference summary. ROUGE-L, on the other hand, measures the longest common subsequence between the two summaries, taking into account also the order of words. All three metrics compute precision, recall, and F1 scores, which are commonly used to evaluate the performance of abstractive summarization models. ROUGE was implemented using the Huggingface's Evaluate[6] library. To be in line with prior work on abstractive summarization, as well as to compare the two variants of the $ROGUE$ metric, the $ROUGE_{RAW}$ was adopted from Straka et al. (2018) for evaluation.

5. Results and Discussion

Evaluation results of the two aforementioned models, together with the simple LEAD and RANDOM baselines models can be seen in Table 5. The presented results show that the mBART model managed to outperform both the LEAD as well as the RANDOM baselines. We note again that this is despite the fact that the pre-trained model has not seen Slovak data in its training. The performance results of mT5 suggest that the model is on par with the RANDOM baseline but it did not manage to outperform the LEAD baseline. This is interesting, as the mT5 model had more training epochs and was pre-trained on Slovak text. At the same time, the mT5 model had tendency to generate shorter but more factual summaries, as exemplified in the sample in Table 9. Additionally, we can observe this phenomena in Table 7, where the average length of references as well as the model outputs computed on the test set are presented. The difference in performance might be due to mBART having 90 million more parameters than mT5. We leave further investigation of this phenomenon to future work.

Model	**Factuality**	**Relevance**
reference	3.01	2.76
mBART	2.94	2.62
mT5	3.35	2.55

Table 6: The factuality and relevance scores of references and outputs from mBART and mT5 models expressed using the Likert, as introduced in Fabbri et al. (2021), on scale from 1 to 5.

Model	**Length**
Reference	139.97
mBART	97.00
mT5	68.57

Table 7: Average output lengths (in characters) of the reference as well as mBART and mT5.

With regards to factuality, the standard approach for a rich-resourced language such as English would be to use a metric such as SummaC or FactCC as an automated evaluation metric. As our work is concerned with Slovak which often lacks data resources necessary for such methods to be viable (e.g. to the best of our knowledge a Slovak-specific NLI dataset does not exist), we instead

[6] https://huggingface.co/docs/evaluate/index

Model	**Precision**	**Recall**	**F1 Score**
LEAD	0.6661	0.6659	0.6649
RANDOM	0.6588	0.6510	0.6539
mBART	0.6192	0.6093	0.6136
mT5	0.6027	0.5765	0.5888

Table 8: The table shows an evaluation of baseline models, LEAD and RANDOM. In the LEAD baseline approach, the summary was constructed by selecting the initial sentence from the test set. Likewise, in the RANDOM baseline approach, the summary was generated by randomly selecting one sentence. The BERTScore for mBART and mT5 models were evaluated on the test set as well.

sampled 100 articles and evaluated their factuality (or consistency) together with the relevance using the methodology described in Fabbri et al. (2021) on a Likert scale from 1 to 5 (higher is better). As can be seen on Table 6, the mT5 model's factuality on this sample is deemed to be higher than that of mBART as well as that of the references. While the mT5's relevance was lowest from the three.
Additionaly, we evaluate the model responses with BERTScore using SlovakBERT, introduced in Pikuliak et al. (2022), as the BERT component. The results can be seen in Table 8. As the results show, the BERTScore F1 is higher for both the LEAD and RANDOM baselines, which suggests that there is a large room for improvement of the finetuned models. However, to the best of our knowledge this is the first time SlovakBERT was used as part of BERTScore and its performance in this regard is understood much less than that of BERT models trained on datasets of rich-resource languages such as English.
Finally, in Table 5, the $ROUGE_{RAW}$ metric was used for evaluation to determine whether the low scores can be potentially attributed to the English-centric text preprocessing. Based on the observed results, it appears that the lack of text preprocessing did not yield a significant impact. We plan to target this issue in our future work and further explore various other evaluation metrics.

6. Conclusions and Future Work

In this paper we introduce `SlovakSum`, a first large scale Slovak news summarization dataset. It consists of more than 200 thousand documents from popular Slovak newswire. Each of the documents in the dataset is composed of a headline, a short abstract consisting of a few sentences, and full source text. Using the presented dataset, we finetuned two supervised abstractive models and reported their performance, as well as the performance of two baseline models using the standard ROUGE metric and $ROUGE_{RAW}$ variant. Additionally, the outputs were evaluated using BERTScore with SlovakBERT component. The code for the reproduction of our dataset and experiments can be found at https://github.com/NaiveNeuron/slovaksum.

reference	Nemecká spolková kancelárka Angela Merkelová reagovala na predpokladané teroristické útoky v Londýne so zdesením a prisľúbila Veľkej Británii solidaritu v boji proti terorizmu. *German Chancellor Angela Merkel reacted with dismay to the presumed terrorist attacks in London and pledged solidarity with the United Kingdom in the fight against terrorism.*
mBART	Nemecká kancelárka Angela Merkelová dnes obvinila Britániu z útokov na Westminsterský most v Londýne, pri ktorých zahynuli najmenej dvaja ľudia. *German Chancellor Angela Merkel today accused Britain of the attacks on Westminster Bridge in London, in which at least two people were killed.*
mT5	Nemecká kancelárka Angela Merkelová dnes upozornila na útoky v Londýne. *German Chancellor Angela Merkel today highlighted the attacks in London.*
reference	V Dúbravke plánujú na jar 2017 otvoriť cintorín pre zvieratá. *In Dúbravka, they plan to open a pet cemetery in the spring of 2017.*
mBART	Cintorín na zvierací pohrebisko v bratislavskej Dúbravke by mohol byť oficiálny už na jar roku 2017. *The cemetery for animal burials in the Bratislava district of Dúbravka could become official as early as the spring of 2017.*
mT5	Nový cintorín pre zvieratá v bratislave by mohol byť otvorený už na jar roku 2017. *The new cemetery for animals in Bratislava could be open as early as spring 2017.*

Table 9: A sample summary generated by mBART and mT5, along with the reference. Note that in the first example, the mBART summary is factually wrong.

Acknowledgements

This research was partially supported by grant APVV-21-0114.

7. Bibliographical References

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Yang Liu. 2019. Fine-tune bert for extractive summarization.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2:159–165.

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference.

Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marian Simko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. 2022. SlovakBERT: Slovak masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7156–7168, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and fine-tuning.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

8. Language Resource References

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints.

George Giannakopoulos, John Conroy, Jeff Kubina, Peter A. Rankel, Elena Lloret, Josef Steinberger, Marina Litvak, and Benoit Favre. 2017. MultiLing 2017 overview. In *Proceedings of the MultiLing 2017 Workshop on Summarization* *and Summary Evaluation Across Source Types and Genres*, pages 1–6, Valencia, Spain. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Milan Straka, Nikita Mediankin, Tom Kocmi, Zdeněk Žabokrtský, Vojtěch Hudeček, and Jan Hajič. 2018. SumeCzech: Large Czech news-based summarization dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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