Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6725–6730
Marseille, 11–16 May 2020

A Summarization Dataset of Slovak News Articles

Marek Šuppa, Jerguš Adamec
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava, Slovakia
marek@suppa.sk, jergus.adamec@gmail.com

Abstract

As a well established NLP task, single-document summarization has seen significant interest in the past few years. However, most of the work has been done on English datasets. This is particularly noticeable in the context of evaluation where the dominant ROUGE metric assumes its input to be written in English. In this paper we aim to address both of these issues by introducing a summarization dataset of articles from a popular Slovak news site and proposing small adaptation to the ROUGE metric that make it better suited for Slovak texts. Several baselines are evaluated on the dataset, including an extractive approach based on the Multilingual version of the BERT architecture. To the best of our knowledge, the presented dataset is the first large-scale news-based summarization dataset for text written in Slovak language. It can be reproduced using the utilities available at `https://github.com/NaiveNeuron/sme-sum`.

Keywords: summarization dataset, document summarization, neural networks, Slovak

1. Introduction

Document summarization is a well-estabilished NLP task which has received considerable attention in the past few years, especially its single-document summarization variant. It is considered to be particularly challenging, as the automatic system needs to understand the text well enough to identify its most important parts, so that it can produce the final summary by either aggregating or rephrasing the previously identified content. Depending on whether they lean towards the former or the latter, such systems can be characterized as *extractive* or *abstractive* (Gambhir and Gupta, 2017).

Much of the recent progress in this area can be attributed to the use of models based on neural networks. While they have shown promising results, particularly in the area of abstractive summarization, they require training corpora on the order of thousands of documents. This has led the authors of (Dernoncourt et al., 2018) to conclude that more large-scale corpora for summarization than those already available are needed. The issue is much more pronounced in the context of summarization of non-English texts, as the data there is small or virtually non-existent. Furthermore, the progress in this area is hindered by the fact that the evaluation metrics used to compare document summarization systems are English-centric (i.e. they depend on English stemmers and stop-words) and cannot therefore be directly applied in other languages.

In this work we try to address both of these issues by introducing a document summarization dataset which consists of Slovak news stories obtained from a prominent Slovak news website. To evaluate its performance, we use a slightly altered version of the ROUGE metric (Lin and Hovy, 2003) which aims to provide a more realistic comparison by utilizing a Slovak stemmer.

2. Related Work

As a task, text summarization has been studied for quite some time, with (Luhn, 1958) attempting to create an "auto-abstract" of technical papers and magazine articles by extracting the sentences with highest significance. Various other extractive approaches have been introduced since then, such as those based on Latent Semantic Analysis (Steinberger and Jezek, 2004), others that use graph-based approaches, such as LexRank (Erkan and Radev, 2004), or TextRank which utilizes PageRank to identify sentences of importance for the final summarization (Mihalcea and Tarau, 2004).

2.1. Neural Summarization Methods

In the recent few years, state-of-the-art results have been achieved using approaches based on neural networks. These model extractive summarization as a sentence classification problem. One of the first examples of such systems include SUMMARUNNER (Nallapati et al., 2017) which encoded the input document with a recurrent neural network. Several other approaches within this paradigm have been subsequently introduced, such as REFRESH that is trained to globally optimize the ROUGE scores with reinforcement learning methods, SUMO (Liu et al., 2019) which sees summarization as tree induction problem, NEUSUM (Zhou et al., 2018) that jointly learns to score and select sentences for the final summarization and BERTSUM (Liu and Lapata, 2019) which exploits the representations provided by pre-trained language models (Devlin et al., 2018).

The abstractive paradigm of single-document summarization has benefited greatly from the introduction of neural matchine translation, since it allows for the task to be formulated as a translation from the source document into the target summary. The first example of the application of this formulation can be found in (Rush et al., 2015) where the authors made use of neural encoder-decoder architecture. Various other alternations have been presented since then, such as pointer-generator networks (See et al., 2017) which are capable of copying words from the source document into the target summary, numerous reinforcement learning-based approaches (Celikyilmaz et al., 2018) (Paulus et al., 2017), models based on convolutional neural networks (Narayan et al., 2018) and those that make use of contextualized representations of pre-trained language

Dataset	no. of documents			mean document length		mean summary length		vocabulary size	
	train	val	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	399,147	81,092
SME	64,001	8,001	8,001	339.09	18.08	23.61	2.16	423,877	110,720

Table 1: A comparison of the presented Slovak dataset with the standard English summarization datasets. In the first multi-column the number of documents in their train/val/test split are shown. The next two describe the mean length (in terms of words and sentences) of the source document and target summary, respectively. Finally, the vocabulary size (of lowercased tokens) for both the document and the summary can be seen in the last multi-column. The values for the English summarization datasets have been taken from (Narayan et al., 2018).

models (Liu and Lapata, 2019).

2.2. Datasets

Historically, the most widely used corpora for single document summarization come from the Document Understanding Conference (Over et al., 2007). Their small size (on the order of hundreds of documents) have made them impractical for training neural network-based models and a need for much larger corpora arose. Recent English summarization approaches are generally tested on the following datasets: **CNN/DailyMail** (Hermann et al., 2015), **NY Times** (Durrett et al., 2016) and **XSum** (Narayan et al., 2018). Since they contain tens of thousands of documents, they are well poised for neural network training regimen. A more comprehensive overview of document summarization corpora can be found in (Dernoncourt et al., 2018).

With regards to non-English text summarization, the most notable corpora are the MultiLing datasets which aim to assist in efforts to improve multilingual summarization. For instance the MultiLing 2015 dataset (Giannakopoulos et al., 2015) contains documents in Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian and Spanish, while the MultiLing 2017 dataset (Giannakopoulos et al., 2017) consists of documents written in 41 languages, including Slovak. The downside of these datasets is thier small size which makes them impractical for use in combination with neural network-based models. However, a large-scale news-based summarization dataset does exist for Czech (Straka et al., 2018), and it is hence closest to the one we present in this work.

2.3. Evaluation Metrics

With regards to evaluation, various metrics have been proposed, including ROUGE (Lin and Hovy, 2003), METEOR (Banerjee and Lavie, 2005) or other LSA-based measures (Steinberger and Ježek, 2012). Despite its shortcomings (Schluter, 2017), ROUGE remains the most widely used automated evaluation metric. In the context of non-English summarization, one of the biggest drawbacks of ROUGE is its English-specificity as it uses English stemmer, English stop words and synonyms. To address this issue, the authors of (Straka et al., 2018) propose an alternative language-agnostic approach called $ROUGE_{RAW}$ that does not use any stemmer and does not consider any stop words or synonyms.

3. The Dataset

The news summarization dataset presented in this work consists of news articles from the web version of a prominent Slovak newspaper SME [1], which was at the time of writing this document the second most popular news website in Slovakia[2]. Following the methodology suggested in (Hermann et al., 2015), we collected over 100 thousand SME.sk articles from the Wayback archive[3]. To ensure that only the articles with full text available were considered, any items that resembled paid content have been removed from further processing. For each article we extracted its headline, short one or two sentence abstract, its actual text, the news category it belongs to (such as Home News, World News, Sport, Travel, Health, Tech, Business and Arts) and the Wayback URL that uniquely identifies it. Based on the unique URL the documents were split into the train/valid/test set in the 80%/10%/10% ratio. Since the abstract is quite short and logically follows the headline, we concatenated the two together in order to create the target summary. Both the source document and the target summary were then tokenized using the `BlingFire` library[4].

A comparison of this dataset with standard English summarization datasets can be seen in Table 1. As we can see, the dataset is split similary to the **XSum**, **NY Times** and **DailyMail** datasets and as Table 2 illustrates, the summary statistics are very similar across all three splits. The length of its documents both in terms of words and sentences is among the smallest ones on the list. This may suggest that it could be applicable in the *extreme summarization* setting which aims to provide a single-sentence answer to the question "What is this article about?" However, since its summary averages to two sentences, direct application in this context would be problematic. Despite its small size, the dataset features a rather rich vocabulary which is most probably due to morphological richness of the Slovak language.

Another way of comparing the presented dataset with the standard English ones can be found in Table 3. The first

[1] Located at https://sme.sk

[2] According to statistics presented on the following ULR `https://web.archive.org/web/20191202122908/https://medialne.etrend.sk/internet-grafy-a-tabulky.html`

[3] `https://archive.org/web/`

[4] `https://github.com/Microsoft/BlingFire`

	Q1	Median	Q3	Mean	SD
train					
document	175	259	402	339.09	323.54
summary	19	22	26.	23.61	6.25
valid					
document	175	264	415	344.99	309.94
summary	19	23	26	23.58	6.36
test					
document	175	257	395	332.25	282.14
summary	19	22	26	23.46	6.14

Table 2: Summary statistics for the length of source documents and target summaries in parts of the presented dataset. The SD column shows the standard deviation while Q1 and Q3 represent the first and third quantile, respectively.

four rows in this table describe the proportion of novel n-grams found in the gold (target) summary. The high number of novel unigrams in the dataset (about 32%) would suggest that it is of a more abstractive nature. It seems that this number is most probably caused by Slovak morphology as when stemming gets applied, the fraction of novel n-grams gets considerably lower (to about 27%).

To further assess the extractiveness or abstractiveness of the **SME** dataset, we evaluate two baseline extractive methods: LEAD and EXT-ORACLE. The LEAD method (Nenkova, 2005), which selects a couple of sentences from the beginning of the source document, can be seen as a strong lower bound on news summarization, as news articles have been traditionally structured such that the most important information is mentioned first. We obtain the metrics for the English datasets from (Narayan et al., 2018), in which the CNN, DailyMail, NY Times and **XSum** have had their LEAD baseline created by extracting the first 3 sentences, first 4 sentences, first 100 words and the first sentence of the source document, respectively. In case of the SME dataset, the first three sentences have been extracted. As the LEAD multi-column in Table 3 shows, the **SME** dataset is much closer to the extractive-leaning datasets like CNN and NY Times than to the abstractive **XSum** dataset.

The EXT-ORACLE method can be seen as an upper bound for extractive approaches, since it creates the candidate summary by selecting the subset of the source document sentences with the highest ROUGE score when evaluated against the target summary. For the **SME** dataset, we select the subset of three sentences as the oracle. The results of evaluating this baseline can be found in the EXT-ORACLE multi-column of Table 3. In spite of the lower values, the **SME** dataset seems to be much closer to the extractive-learning datasets than to the abstractive **XSum** with regards to this baseline as well.

Taking the results of both the LEAD and EXT-ORACLE together, we conclude that the **SME** dataset is much more extractive than abstractive and we therefore focus on extractive models in our experiments.

4. Experiments and Evaluation

4.1. Baselines

To complement the aforementioned LEAD and EXT-ORACLE baselines, we also introduce a RANDOM baseline which simply selects the desired number of sentences from the source document randomly. In contrast to LEAD, the RANDOM baseline can be viewed as a weak lower bound on the presented summarization task.

Furthermore, we also use the TextRank model (Mihalcea and Tarau, 2004), a standard unsupervised approach for document summarization. TextRank represents the sentences in the text as a graph with edge values representing the similarity between sentences, and then uses PageRank to identify the most important ones. In particular, we used the `summa textrank` package[5] which includes optimizations from (Barrios et al., 2016). The aforementioned package was slightly altered to use a list of Slovak stop words and Slovak stemmer.

4.2. BERT-based Extractive Model

To provide a strong baseline for the presented dataset, we use the model introduced in (Liu and Lapata, 2019) with small alternations, to accommodate for the fact that in our case the input data is written in a different language and script.

The model introduced in (Liu and Lapata, 2019) makes use of the Bidirectional Encoder Representations from Transformers (abbreviated as BERT, introduced in (Devlin et al., 2018)), a language representation model trained on large corpora of unannotated text. BERT generalizes the idea of distributed word representations by making them context-specific, as opposed to type-specific. Models of this type are generally trained with a next sentence prediction and masked language learning objectives.

Inspired by the promising results reported in (Pires et al., 2019), we alter the model to use Multilingual BERT (M-BERT) – a variant of BERT pre-trained on a corpora of the top 104 most voluminous Wikipedias. Its aim is to provide language-independent representations, as no specific information denoting the particular input language is provided during training. To this end, the model also features a word piece vocabulary that is shared across all considered languages.

Specifically, our model uses the BERT-Base, Multilingual Cased variant [6] which assumes no input normalization (such as lower-casing, stripping of diacritics marks or Unicode normalization). Similarly to the standard English BERT-Base models, the M-BERT consists of 12 Transformer layers (Vaswani et al., 2017), each of them with 768 hidden units, resulting in a model with about 110 million parameters. Prior to being used as input to the model, all of the texts were tokenized using the BERT tokenizer.

The model was implemented by updating the code graciously provided by the authors of (Liu and Lapata, 2019)[7], which is based on PyTorch (Paszke et al., 2019), OpenNMT

[5] `https://github.com/summanlp/textrank`

[6] This model is also known under the name `bert-base-multilingual-cased`

[7] `https://github.com/nlpyang/PreSumm`

Dataset	percent of novel n-grams in gold summary				LEAD			EXT-ORACLE		
	unigram	bigram	trigram	4-gram	R-1	R-2	R-L	R-1	R-2	R-L
CNN	16.75	54.33	72.42	80.37	29.15	11.13	25.95	50.38	28.55	46.58
DailyMail	17.03	53.78	72.14	80.28	40.68	18.36	37.25	55.12	30.55	51.24
NY Times	22.64	55.59	71.93	80.16	31.85	15.86	23.75	52.08	31.59	46.72
XSum	**35.76**	**83.45**	**95.50**	**98.49**	**16.30**	**1.61**	**11.95**	**29.79**	**8.81**	**22.65**
SME	**32.40**	**62.06**	**73.61**	**79.59**	**29.67**	14.04	25.41	41.28	28.09	39.18
SME + stem	27.02	60.00	72.83	79.20	29.27	**14.66**	**26.68**	**42.77**	**28.96**	**40.41**

Table 3: An empirical comparison of the **CNN/DailyMail**, **NY Times**, **XSum** and **SME** datasets with regards to their extractive/abstractive nature. The first four columns denote the percentage of novel n-grams in the target summaries. The remaining columns show the ROUGE-1, ROUGE-2 and ROUGE-L scores for both the LEAD and EXT-ORACLE baselines. Metrics for the English datasets are reproduced from (Narayan et al., 2018).

(Klein et al., 2017) and HuggingFace's PyTorch Transformers (Wolf et al., 2019).

Similarly to the process the original authors described, we trained the model for 50,000 steps on a single GPU (GTX 1080). Model checkpoints were saved after each 1,000 steps and gradient was accumulated on every second step. Since the considered dataset contains abstractive target summaries, a greedy algorithm similar to that introduced in (Nallapati et al., 2017) was used to provide extractive targets. It selects sentences which maximize the ROUGE-2 score when compared against the abstractive summaries.

To infer which sentences to extract from a new document, the model predicts a score for each sentence in the source documents and the top 3 sentences are then selected as the extracted summary. In order to reduce redundancy, the trigram blocking procedure[8] described in (Paulus et al., 2017) is used.

4.3. Evaluation

To automatically evaluate the predicted summaries we use the ROUGE metric (Lin and Hovy, 2003). In particular, the ROUGE-1 and ROUGE-2 scores represent the unigram and bigram overlap and ROUGE-L represents the longest common subsequence. The first two metrics can be seen as a proxy measure for informativeness while the last metric can be taken to represent fluency.

As we mentioned in Section 1., one of the issues with ROUGE applied to non-English texts is that it makes use of English-specific components, such as an English stemmer for instance. To remedy this problem, (Straka et al., 2018) ditch the stemming part altogether, resulting in a language-agnostic approach called $ROUGE_{RAW}$. In our experiments, however, this approach proved to be problematic as it sometimes meant that the automatic evaluation reported zero score, even if the summary was in fact of high quality. The authors also note this fact themselves at the end of the *Examples* section of the aforementioned study.

We try a different approach, in which both the system and reference summaries are passed through a Slovak stemmer before the ROGUE score is computed. To implement this approach, we update a Python implementation of the ROUGE score evaluation package `py-rouge`[9] to work with a custom Python-based stemmer. To obtain ROUGE score, as well as in any other case when Slovak stemming was necessary or appropriate, we used the `stemmsk` package [10]. This package adapts the Czech stemmer described in (Dolamic and Savoy, 2009) and we use the "light" version throughout this study.

5. Results and Discussion

The results of evaluation of the aforementioned models can be seen in Table 4 In the interest of succinctness, the ROUGE F1 scores are reported.

Model	R-1	R-2	R-L
RANDOM	21.62	8.74	18.43
LEAD	29.27	14.66	26.68
EXT-ORACLE	**42.77**	**28.96**	**40.41**
TEXTRANK	22.28	9.00	19.69
M-BERT	**29.38**	**14.69**	**26.79**

Table 4: The ROUGE F1 scores of various baselines and extractive models reported on the **SME** test set.

We observe that while the unsupervised TEXTRANK managed to outperform the weak lower bound RANDOM, it did so only by a slight margin – as much as 0.26 ROUGE points in the case of ROUGE-2. Similarly, the M-BERT model managed to achieve better results than the LEAD baseline but the differences in this case seem negligible – only 0.03 ROUGE points in case of ROUGE-2. Considering the values reported for the EXT-ORACLE baseline, it is clear that the presented models leave a substantial room for improvement of extractive methods on this dataset in the future.

We also need to critically note that the dataset may suffer from some of the issues described in (Kryściński et al., 2019). Namely, it is possible that the information contained in the dataset leaves the task under constrained and too ambiguous to be solved with end-to-end models, such as those whose results were presented above. Given the fact that a comparable dataset for Slovak language does not exist to the best of our knowledge, we still consider it a valuable contribution to the research community.

[8] A sentence is not selected for the summary if there exists a trigram overlap between this sentence and the existing summary.

[9] https://github.com/Diego999/py-rouge

[10] https://github.com/mrshu/stemm-sk/

6. Conclusions and Future Work

In this work we introduce a Slovak news-based summarization dataset. It consists of tens of thousands of articles which contain the source document and a concatenation of the document's headline and its short abstract that constitute the target summary. We evaluate various baselines as well as supervised and unsupervised extraction methods and report their results using an adaptation of the ROUGE metric for Slovak texts. The code used for the experiments as well as the utilities necessary for producing the dataset can be found at `https://github.com/NaiveNeuron/sme-sum`.

The existence of large scale document summarization datasets in various languages provides new research opportunities, especially in fields like transfer learning. Despite its shortcomings, we hope it may serve as a test bed for future studies.

7. Bibliographical References

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Barrios, F., López, F., Argerich, L., and Wachenchauzer, R. (2016). Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606.

Celikyilmaz, A., Bosselut, A., He, X., and Choi, Y. (2018). Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675, New Orleans, Louisiana, June. Association for Computational Linguistics.

Dernoncourt, F., Ghassemi, M., and Chang, W. (2018). A repository of corpora for summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dolamic, L. and Savoy, J. (2009). Indexing and stemming approaches for the czech language. *Information Processing & Management*, 45(6):714–720.

Durrett, G., Berg-Kirkpatrick, T., and Klein, D. (2016). Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany, August. Association for Computational Linguistics.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Gambhir, M. and Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.

Giannakopoulos, G., Kubina, J., Conroy, J., Steinberger, J., Favre, B., Kabadjov, M., Kruschwitz, U., and Poesio, M. (2015). Multiling 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274.

Giannakopoulos, G., Conroy, J., Kubina, J., Rankel, P. A., Lloret, E., Steinberger, J., Litvak, M., and Favre, B. (2017). Multiling 2017 overview. In *Proceedings of the MultiLing 2017 workshop on summarization and summary evaluation across source types and genres*, pages 1–6.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Lin, C.-Y. and Hovy, E. (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.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Liu, Y., Titov, I., and Lapata, M. (2019). Single document summarization as tree induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1745–1755, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Nallapati, R., Zhai, F., and Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference.

Over, P., Dang, H., and Harman, D. (2007). Duc in context. *Information Processing & Management*, 43(6):1506–1520.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, September. Association for Computational Linguistics.

Schluter, N. (2017). The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July. Association for Computational Linguistics.

Steinberger, J. and Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4:93–100.

Steinberger, J. and Ježek, K. (2012). Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.

Straka, M., Mediankin, N., Kocmi, T., Žabokrtskỳ, Z., Hudeček, V., and Hajic, J. (2018). Sumeczech: Large czech news-based summarization dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M., and Zhao, T. (2018). Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia, July. Association for Computational Linguistics.