Error Analysis of using BART for Multi-Document Summarization: A Study for English and German Language

Timo Johner[1], **Abhik Jana**[1], and **Chris Biemann**[1]

[1] Language Technology Group, Dept. of Informatics, Universität Hamburg, Germany
me@timojohner.de, jana@informatik.uni-hamburg.de
biemann@informatik.uni-hamburg.de

Abstract

Recent research using pre-trained language models for multi-document summarization tasks lacks a deep investigation of potential erroneous cases and their possible application in other languages. In this work, we apply a pre-trained language model (BART) for multi-document summarization (MDS) task, both with fine-tuning and without fine-tuning. We use two English datasets and one German dataset for this study. First, we reproduce the multi-document summaries for the English language by following one of the recent studies. Next, we show the applicability of the model to the German language by achieving state-of-the-art performance on German MDS. We perform an in-depth error analysis of the followed approach for both languages, which leads us to identify the most notable errors, from made-up facts to topic delimitation. Lastly, we quantify the amount of extractiveness.

1 Introduction

Nowadays, we are confronted with an enormous amount of information through news, mails, social media, etc., which are difficult to absorb for a human being in one go. Hence, there is a pressing need to compress and comprehend this information. Capturing salient details from multiple sources to produce an abridged version is described as Multi-Document Summarization (MDS) (Nenkova and McKeown, 2011) and can be carried out in both an abstractive or extractive manner. MDS has recently become one of the most interesting research topics in the field of Natural Language Processing (NLP). As per the literature, whilst the state-of-the-art models (Gehrmann et al., 2018; Liu et al., 2018) heavily rely on large datasets, recent advances with pre-trained language model systems (Ziegler et al., 2019; Raffel et al., 2020; Lewis et al., 2020) have shown great potential for the summarization task. While there have been studies to gradually improve the performance of MDS for the English language, MDS for other languages has rarely been attempted. There has also been a lack of in-depth error analysis for the MDS task. In this study, we attempt to analyze and address these issues.

Our main contributions are the following: Firstly, we reproduce recent pre-trained and fine-tuned results for multi-document summarization with the BART model, introduced by Lewis et al. (2020), on two English datasets. Further, we adapt the model for the German language and achieve state-of-the-art performance for the German MDS task, beating the most competitive baseline by a margin of 3.48-8.67%. Secondly, we perform an analysis on the erroneous cases for both languages where we point out general errors and cross-lingual error similarities regarding factfulness and topic delimitation. Additionally, we also investigate the extractiveness of the generated summaries.

2 Related Work

Early approaches on extractive MDS apply term frequency-inverse document frequency (TF-IDF) (McKeown et al., 1999; Goldstein et al., 2000; Radev et al., 2000). Later, Conroy et al. (2006) and Shen and Li (2010), attempt the MDS task with a topic and set-based methodology, respectively. Initial attempts for abstractive multi-document summarization are made by McKeown and Radev (1995) and Radev and McKeown (1998). Barzilay and McKeown (2005) use sentence-fusion for text generation to create summaries across different documents. Haghighi and Vanderwende (2009) build a model based on word frequency and Latent Dirichlet Allocation (LDA) for MDS whereas phrase selection and merging approaches have also been tried (Bing et al., 2015) for the same.

In recent years, neural network architecture is being adapted for several NLP tasks, especially with the approach of using encoder-decoder architecture. Here, relevant work includes Rush et al. (2015), who propose an attention model for combining extractive and abstractive methods, which is supplemented with document-wide contextualization by Cheng and Lapata (2016) and Nallapati et al. (2016). In a different direction, several graph-based approaches are explored as well (Tan et al., 2017; Yasunaga et al., 2017). Liu et al. (2018) show the feasibility of using Wikipedia as an MDS dataset whereas Fabbri et al. (2019) apply a pointer-generator network with a transformer model complemented with Maximal Marginal Relevance (MMR). Li et al. (2020) explores graph representation and proposes to leverage graphs for abstractive MDS.

Most recently, fine-tuning pre-trained language models have gained a lot of attention for NLP tasks. For summarization, one such work by Raffel et al. (2020) attempted to explore fine-tuning, whereas, in another work, Liu and Lapata (2019) fine-tune BERT for summarization. Later, Hokamp et al. (2020) adapt and fine-tune BART on MDS. Approaches regarding a systematic error analysis of those models were introduced by Huang et al. (2020) who compared BART to other abstractive and extractive methods.

In another direction, attempts have also been made for single-document summarization for non-English text. For instance, single-document summarization of text in German language was done by Parida and Motlicek (2019) who utilized transformer models for abstractive summarization on two datasets — SwissText 2019[1] and Common Crawl[2]. Evaluation of summarization models to non-English data was done by Tauchmann and Mieskes (2020) who applied an automatic evaluation paradigm on the German heterogeneous dataset DBS (Benikova et al., 2016). Since our main focus is on multi-document summarization, we do not explore the literature of single-document summarization extensively.

3 Datasets

For our experiments we use three datasets that exhibit extractive characteristics: two English datasets — CNN/DM (Hermann et al., 2015), Multi-News (Fabbri et al., 2019) and one German dataset — auto-hMDS (Zopf, 2018).

CNN/DailyMail This dataset is an English single-document summarization (SDS) news dataset consisting of 311,971 news articles with an average length of ∼800 words from the CNN and DailyMail websites including abstractive summaries.

Multi-News The Multi-News dataset is an English MDS news dataset consisting of 56,216 summaries and over 250,000 sources with an average of ∼2,100 words from 1,500 different sites. The summaries are linked to 2-10 human-written source documents retrieved from `https://www.newser.com/`.

auto-*h*MDS This is the largest dataset for multi-document summarization in German language with 2,210 summaries and 10,454 source documents, and diverse in nature. The dataset is created by selecting available summaries from Wikipedia and search for corresponding source documents on the internet. On an average, a summary is linked to 4.73 source documents.

4 Methodology

We consider the state-of-the-art BART model (Lewis et al., 2020) for multi-document summarization (MDS) task. First, we use only pre-trained BART for the task, and next, we fine-tune the pre-trained BART model using each of the three datasets separately and analyze the performances. The details about the BART model are described below.

Description of BART model BART (Lewis et al., 2020) generalizes the concepts of bidirectional encoders from BERT (Devlin et al., 2019) and autoregressive decoders from GPT-2 (Radford et al., 2019). The model is trained with text corrupted through an arbitrary noising function and a sequence-to-sequence model that learns to reconstruct the original text. The encoder reads the sequential input e.g. a document to summarize while the decoder generates the outputs autoregressively. Both layers are connected by cross-attention where each decoder layer focuses on specific aspects over the final state of the encoder output creating sequences, closely connected to the initial input. The bidirectional encoder architecture takes all previous and subsequent tokens into account for predicting a masked token. In text generation, BERT without any modification loses its strength

[1] `https://www.swisstext.org/`

[2] `http://commoncrawl.org/`

of bi-directionalism and becomes directional towards past words, as following words have yet to be generated. Here BART adopts the architecture of GPT-2 to predict future words only by utilizing previous words. The advantage of BART therefore is the combination of contextual embeddings from BERT and text generation from GPT-2. Transformation, as described in Lewis et al. (2020), can be implemented through token masking, token deletion, text infilling, sentence permutation, or document rotation.

Note that, in the work done by Lewis et al. (2020), authors apply the BART model only on single-document summarization (SDS) task, not on the multi-document variant of the summarization task. Therefore, to adapt the BART model for the MDS task, we follow the approach prescribed by Lebanoff et al. (2018), where authors reuse the existing SDS model for MDS by merging multiple-input to single-input. On the other hand, the issue of redundant and overlapping information is one major point to be taken care of for any summarization task, especially for MDS tasks. For that purpose, we rely on the n-gram blocking approach following the work done by Paulus et al. (2017).

5 Experimental Results and Error Analysis

For our experiments we make use of the pre-trained BART model[3] and fine-tune the model on the three datasets and compare the performance with competitive baselines.

Method	**R-1**	**R-2**	**R-L**
Lead-3 (Liu and Lapata)	40.42	17.62	36.67
BERTSUMABS (Liu and Lapata)	41.72	19.39	38.76
BERTSUMEXTABS (Liu and Lapata)	42.13	**19.60**	**39.18**
BART pre-trained	25.98	11.26	17.50
BART fine-tuned	**42.21**	19.10	35.38

Table 1: Performance of the BART pre-trained and fine-tuned models along with most competitive baselines (Liu and Lapata, 2019) on CNN/DM dataset.

5.1 Comparative Evaluation

We split each dataset into training (80%), validation (10%) and test (10%) set. In our experimental setup, we use the beam size of 4, n-gram size of 3 and use the Adam optimizer (Kingma and Ba, 2014)[4]. To evaluate the the model generated summaries, we use the variants (**R-1, R-2, R-L**) of ROUGE metric (Lin, 2004) as required for the comparison with the baseline models.

Method	**R-1**	**R-2**	**R-L**
Hi-Map (Fabbri et al.)	40.08	14.90	19.70
BART DYNE-1 (Hokamp et al.)	**43.90**	**15.80**	**22.20**
BART DYNE-5 (Hokamp et al.)	43.20	13.60	20.40
BART pre-trained	30.67	10.05	16.99
BART fine-tuned	40.58	15.50	21.73

Table 2: Performance of the pre-trained and fine-tuned BART model along with baseline models on Multi-News Dataset.

	100 words		200 words	
Method	**R-1**	**R-2**	**R-1**	**R-2**
Random (Zopf)	18.57	1.85	25.53	3.25
Lead (Zopf)	12.29	2.61	10.56	2.28
Top-5 Sentences	21.71	4.28	19.61	3.87
LexRank	29.76	6.58	23.81	5.61
BART pre-trained	28.48	8.79	20.84	6.02
BART fine-tuned	**38.43**	**12.93**	**30.24**	**9.09**

Table 3: Performance of the BART pre-trained and fine-tuned model along with baseline models on auto-hMDS dataset.

Summary (gold)
A South Carolina man says he spent 66 days alone at sea before being rescued . Other sole survivor stories include a Japanese man washed away by a tsunami . An El Salvador man says he drifted from Mexico to Marshall Islands over a year .
Summary (generated) R-1 = 12.50 R-2 = 2.53 R-L = 8.75
Sailors can't bank on technology or the proximity of a nearby city, town, or boat. In order to survive, they can rely on ingenuity, resourcefulness and luck. Jose Salvador Alvarenga says his journey began in Paredon Viejo, a port on Mexico's Pacific coast, in late 2012. He says he drank rainwater and when there wasn't any available, his own urine. Louis Jordan says he used laundry to trap and scoop up fish, rigged a makeshift mast, and sail.

Table 4: Example of missing facts within the CNN/DM dataset.

Table 1 shows the performance of the pre-trained and fine-tuned BART model on the CNN/DM dataset, along with the performance of the baseline models. We see, that the fine-tuned BART model produces comparable performance with the baselines. On the other hand, Table 2 shows the results of the pre-trained and fine-tuned BART model on the Multi-News dataset. We observe the fine-tuned model outperform the Hi-Map (Fabbri et al., 2019) model, whereas it produces comparable performance with BART-DYNE (Hokamp et al., 2020). Note that, the fine-tuned BART model considers all source documents for the MDS task whereas the model by Hokamp et al. (2020) only takes one (DYNE-1) or five source documents (DYNE-5) into account, which otherwise simplifies the task. Table 3 shows the results on the German auto-hMDS dataset of pre-trained and fine-tuned BART models in comparison to baselines proposed by Zopf (2018). We prepare two baseline

[3] `https://github.com/pytorch/fairseq/tree/master/examples/bart`

[4] Default settings $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a learning rate of $3e-05$

models as well. The first one is trivial by extracting 'Top-5 Sentences' based on the frequency of occurring words and the second one by following the LexRank (Erkan and Radev, 2004) approach. We see that the fine-tuned BART model outperforms all the baseline models by a significant margin, producing a state-of-the-art performance for the German MDS task[5].

5.2 Error Analysis

Even though the BART model produces satisfactory performance for multi-document summarization for both languages, there is still scope for improvement. Hence, we investigate cases further, where even the fine-tuned BART model goes wrong. We perform this analysis for both English and German languages. To start with, we observe some interesting cases for which the model does not generate the desired gold summary due to the fact that some information in the gold summary is actually not present in any of the source documents. Table 4 represents one such interesting error case obtained from the CNN/DM dataset.

Table 5 shows one example from another frequently occurring genre of erroneous cases, for the Multi-News dataset (at top) where the model generated summary is very meaningful and comprehensive but makes up new facts such as the death of Bob Dylan (color-coded in orange). We perform a manual survey on the randomly selected model-generated summaries and observe at least 4 out of 50 summaries which include made-up facts in an otherwise coherent summary. This very pattern can also be seen while experimenting with the German auto-hMDS dataset (Table 5, at the bottom), where the place and date of birth are made-up facts. This is misleading as wrong facts are embedded in a reasonable and correct context, making them especially hard to spot.

Another genre of erroneous summaries, which we detect while experimenting with the German auto-hMDS dataset, comes from lacking clear contextualization and topic delimitation. Table 6 presents one such example, where the model should summarize information about the 'Palace of Westminster', but as the source document includes references to related buildings, the model lost attention and mixed up information about the 'Palace of Westminster' and 'Westminster Abbey' (in orange) in one summary.

Summary (model generated) R-1 = 67,59, R-2 = 29.91, R-L = 31.41
[...] The former James Bond star, 65, who was trained as a commercial artist and worked as an illustrator, just auctioned off one of his paintings for $1.4 million, depicting the singer, who died in 2013. Other auction highlights included a Pierce Brosnam original painting, which sold for

Summary (model generated) R-1 = 55.88, R-2 = 11.94, R-L = 30.88
Andrew Johnson (* 29. Dezember 1808 in Raleigh (North Carolina, USA; † 15. April 1865 in Greeneville, Tennessee) war der dritte Vizepraesident der Vereinigten Staaten, der durch den Tod seines Vorgaengers ins Amt kam und der erste nach einem Attentat. Als Hauptaufgabe seiner Praesidentschaft galt die sogenannte Reconstruction, der Wiederaufbau [...]

Table 5: Examples of summaries showing wrong facts while experimenting with the Multi-News (Top) and auto-hMDS (Bottom) datasets.

Summary (model generated) R-1 = 75,81, R-2 = 31.14, R-L = 32.53
Die Westminster Abbey ist die Kroenungskirche der bristischen Monarchen seit Wilhelm dem Eroberer im 11. Jahrhundert. Erbaut wurde die Westminster Abbey zwischen 1045 und 1065 auf dem Kloster Kloster der Themse an den damals noch sumpfigen Ufern der themse errichtet. Bis zum Jahr 1529 diente der Palast den britischen Koenigen als Residenz. Heute ist der neugotische Palast vor allem als Houses of Parliament bekannt.

Table 6: Example of summary showing wrong contextualization and topic extraction while experimenting with auto-hMDS dataset.

5.3 Analysis of Extractiveness of Summaries

After analyzing and pointing out the erroneous cases, we further investigate the nature of model-generated summaries along with the gold summaries of each dataset, in terms of extractiveness. Even though according to one of the recent studies (Lewis et al., 2020), the BART model output is "highly abstractive, with few phrases copied from the input", our findings are contrary with summaries mainly built from extractive fragments or even whole paragraphs. To investigate quantitatively, we measure the extractiveness by using the method of extractive coverage and extractive density, introduced by Grusky et al. (2018)[6].

From Figure 1, we can see that the model-generated summaries from fine-tuned BART are much more extractive than their gold counterparts with an average extractive coverage over 94%. While the gold summaries are already much more extractive, BART generated summaries increase extractiveness further. The figure also discloses the difference between the German auto-hMDS dataset and the English datasets. The average extractive density of the gold summaries from the German

[5] Note that, we do not report the R-L score in Table 3, as R-L scores are not reported for the baseline models used for comparison in the previous works.

[6] Extractive density measures how well a sequence of a summary is made of extractions from the source, while extractive coverage measures, how much the summary is a derivative of the source taking into account individual words. Note that, many individual words could result in a high coverage, but a high density can only be achieved with long extractive fragments.

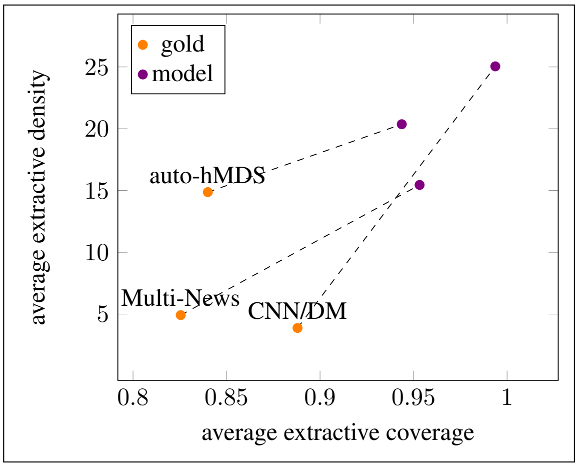

Figure 1: Comparison of extractiveness of gold summaries and model-generated summaries.

auto-hMDS shows that the summaries are mainly built from long extractive fragments, much more than the English gold standard summaries.

6 Conclusion

In this paper, we investigated the performance of one of the most recent pre-trained language models namely BART, for multi-document summarization tasks in English and German language. For the first time ever, we attempted fine-tuning BART for German language multi-document summarization and achieved state-of-the-art performance. We further analyzed the erroneous cases for both English and German language and attempted to find a set of patterns where BART went wrong. The insights obtained via this error analysis give rise to devise more sophisticated methods for the task of multi-document summarization addressing these errors, of which the most severe is the hallucination of facts.

Our code and data repository is available publicly[7].

References

Regina Barzilay and Kathleen R McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Darina Benikova, Margot Mieskes, Christian M. Meyer, and Iryna Gurevych. 2016. Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1039–1050, Osaka, Japan.

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca J Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1587–1597, Beijing, China.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany.

John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 152–159, Sydney, Australia.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, USA.

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

Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium.

Jade Goldstein, Vibhu O Mittal, Jaime G Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*, pages 40–48, Seattle, Washington, USA.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana, USA.

[7] `https://github.com/uhh-lt/multi-summ-german`

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado, USA.

Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1693–1701, Montreal, Quebec, Canada.

Chris Hokamp, Demian Gholipour Ghalandari, Nghia The Pham, and John Glover. 2020. Dyne: Dynamic ensemble decoding for multi-document summarization. *arXiv preprint arXiv:2006.08748*.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6232–6243, Online.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Łukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations (ICLR)*, Vancouver, British Columbia, Canada.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731, Hong Kong, China.

Kathleen McKeown and Dragomir R Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, Seattle, Washington, USA.

Kathleen R McKeown, Judith L Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: progress and prospects. In *Proceedings of the sixteenth AAAI Conference on Artificial Intelligence*, pages 453–460, Orlando, Florida, USA.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany.

Ani Nenkova and Kathleen McKeown. 2011. *Automatic summarization*. Now Publishers Inc.

Shantipriya Parida and Petr Motlicek. 2019. Abstract text summarization: A low resource challenge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5994–5998, Hong Kong, China.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Dragomir Radev and Kathleen McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Dragomir R Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*, pages 21–30, Seattle, Washington, USA.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *Journal of Machine Learning Research*, 21:1–67.

Alexander M Rush, Sumit Chopra, and Jason Weston. 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.

Chao Shen and Tao Li. 2010. Multi-document summarization via the minimum dominating set. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 984–992, Beijing, China.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1181, Vancouver, British Columbia, Canada.

Christopher Tauchmann and Margot Mieskes. 2020. Language agnostic automatic summarization evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6656–6662, Marseille, France.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462, Vancouver, British Columbia, Canada.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Markus Zopf. 2018. Auto-hMDS: Automatic construction of a large heterogeneous multilingual multi-document summarization corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 3228–3233, Miyazaki, Japan.