ExtraPhraseRank: A Length-Controlled Data Augmentation Strategy for Unsupervised Abstractive Summarization

Stanford CS224N Custom Project

Daniel Fein
Department of Computer Science
Stanford University
drfein@stanford.edu

Ruben Cuevas
Department of Computer Science
Stanford University
ruben.cuevas@stanford.edu

Abstract

The emergence of large pre-trained language models has redefined the state of the art for abstractive summarization. Unfortunately, the lack of parallel data for this task is a hindrance to the usefulness of these models. Further, existing methods of unsupervised summarization often lack control of their output length. We solve these problems by presenting ExtraPhraseRank, a data augmentation strategy for the efficient, length-controlled generation of synthetic summaries. The algorithm has two major steps: it first invokes the TextRank algorithm [1] to extract the most important sentences in a text. It then back-translates the most important sentences (by translating them to German, and back to English) in order to increase the diversity of words and phrases. We test this method on a popular abstractive summarization dataset, and then use the generated summaries to fine-tune a BART model [2] for the document summarization task. We find that this strategy produces modest improvements in ROUGE scores above our base-lines, but that it is ultimately not as powerful as fine-tuning even a smaller number of human-written summaries. Finally, we test the practicality of our algorithm on a multi-document summarization task with no existing labels (namely, summarizing collecions of amazon reviews). We find that, though promising in the way of controlling the output length of summaries generated, the algorithm does not lead to summaries significantly more useful than a vanilla BART model.

1 Key Information to include

- Mentor: Manan Rai
- External Collaborators (if you have any): N/A
- Sharing project: No

2 Introduction

Document summarization is an important task with many real world applications. The two primary modes of summarization are extractive and abstractive summarization. Extractive summarization involves selecting a substring from the source document that best represents the content of that source, while abstractive summarization is a language generation task in which a novel summary sentence is constructed that compresses the information from the source document. Despite the promise of neural summarization techniques in recent years, it is not yet widely applicable to real world problems due to a lack of document-summary pairs. While there are many data augmentation techniques that seek to remedy this problem by modifying existing document-summary pairs to increase training data, fewer methods of fully unsupervised summarization exist. Of those that do, most are purely extractive [3]. This problem is exacerbated by the task of multi-document summarization, where the source text consists of multiple (often redundant) pieces of similar text; datasets for this task are even more rare. This makes sense, as summaries cost time and money to write in large quantities.

In recent years, massive pre-trained language models such as GPT-3 and BERT have made significant progress across NLP tasks. Fine-tuning these models is particularly powerful, as it enables models to be trained on less data while still producing good results. However, the task of summarization, and specifically multi-document summarization, has yet to reap the benefits of fine-tuning due to the lack of data described above.

Another documented challenge in unsupervised abstractive summarization is the length limit problem [4]. It is often essential to the usefulness of summaries that they are of a specific length: too long, and they provide no benefit over the original source text, too short, and they omit key information. Unsupervised algorithms must be able to control the length of output sentences in order for them to be widely useful. This is a key shortcoming of a data augmentation algorithm called ExtraPhrase, published earlier this year [5]. This algorithm demonstrated success at creating data that improved a transformer-based model's ability to generate summaries. However, it falls short of being widely useful, as its dependency parsing-based system for extracting information from source texts does not allow for control of output summary length.

We solve this problem, by introducing ExtraPhraseRank, a data augmentation method that generates pseudo-summaries from documents with fine control over word count. Moreover, with control of word length, we propose that ExtraPhraseRank enables the finetuning of pre-training large language models without document-summary pairs. We achieve some success in demonstrating this, but ultimately find that fine-tuning on even a small number of gold document-summary pairs can generate higher ROUGE scores than fine-tuning on purely synthetic examples. Lastly, we demonstrate the practicality of this algorithm by producing summaries of user reviews by fine-tuning BART on ExtraPhraseRank-generated pseudo-summaries.

3 Related Work

3.1 Abstractive Summarization

Previous attempts have been made to create abstractive summaries using a transformer language model. These abstractive models have historically been sound in producing diverse outputs and rewording key phrases. Attention based encoder/decoders have been researched heavily [6] and applied to summarization tasks specifically. Some of these modify the approaches to better summarize a given text. For example, one approach by Subramanian et al. chooses a few important sentences and then summarizes only those using the transformer method, outperforming the traditional seq-to-seq transformer [7]. As another example method, other researchers have looked into using social context such as comments of online posts to summarize the posts themselves [8].

3.2 Data Augmentation

A key insight we gained from previous work was the lack of document-summary pairs in the data. The XSum dataset was in part created to try to fix this, as it contains over 200,000 summaries of articles paired with the originals [9]. As mentioned earlier, other abstractive summarization methods have been primarily taking parts of a document and summarizing specific sentences, but there is a lack of data and document-summary pairs while doing this task. Previous attempts have been made at data augmentation for seq-to-seq methods, and many of them have used backtranslation to construct this pseudo training data [5]. However, it is to be noted that only backtranslation for a summarization task is difficult due to the fact that on one part of backtranslation, a source is expected to be generated from a summary. Such pseudo data generation methods have been outperformed by methods such as ExtraPhrases, which has also been shown to be more cost-effective [5].

4 Approach

4.1 Data Augmentation Method

To generate synthetic summaries for fine-tuning, we employ a two-step approach. As is done in ExtraPhrase [5], we first perform an extractive summarization technique, and then perform backtranslation on these summaries (translating from English to German, and then back to English) in order to obtain diverse abstractive pseudo-summaries for each document. Instead of using extractive summarization techniques that work on the level of sentence dependencies, as in ExtraPhrase, we use the textRank algorithm [1] in order to obtain sentences of a specified length. We implement this operation by using the textRank implementation at https://github.com/summanlp/textrank [10], as well as the Helsinki Opus neural translation model. When [11]. The rest of the code to create synthetic data, as well as to fine-tune and test BART, was written by us.

Figure 1: The ExtraPhraseRank Pipeline

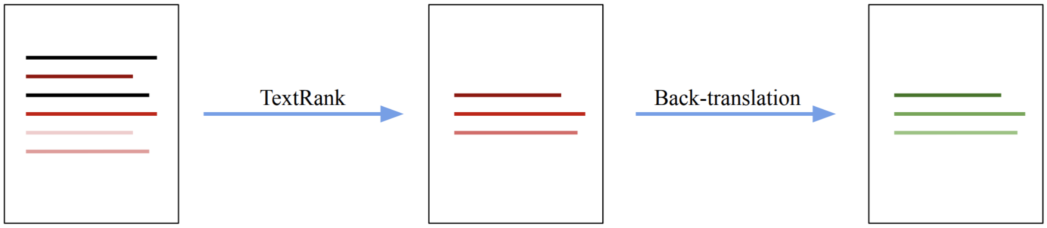
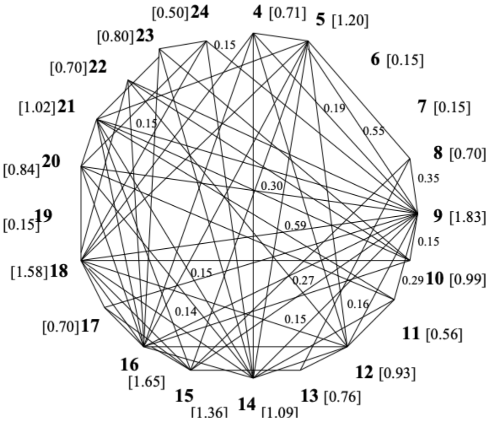


Image created by authors.

4.2 The TextRank Algorithm

The first component of our data augmentation algorithm is extractive summarization. We use the well-known TextRank algorithm [1] to carry this out. This has two primary advantages. First, it produces reliable, extractive summaries in both single and multi-document contexts. Second, it is an efficient and low-cost algorithm that could theoretically be used on a very large corpus of documents to generate improved summaries. TextRank is an unsupervised extractive summarization algorithm based on the PageRank algorithm [12]. It functions by first adding each sentence in a text as a node into a fully connected, undirected graph. It then weights each edge by the similarity of their contents. The implementation we used, for efficiency reasons, uses a probabilistic model called BM25 to calculate these similarities [13]. Once this graph is constructed, the PageRank algorithm ranks the sentences in terms of importance (based on their similarity to other sentences). The top sentences are then returned based on the number of words specified.

Figure 2: Sample TextRank Graph



Each node in this graph represents a sentence, and each edge represents similarity between sentences. Lastly, the numbers in brackets represent the calculated importance of the sentences based on their similarity to other sentences. Image from [1].

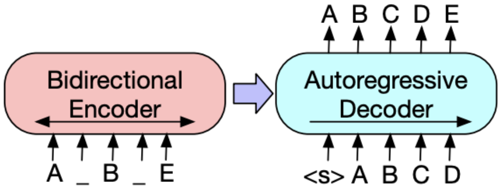
4.3 Back-Translation

Back-translation is a popular data augmentation paradigm that relies on neural machine translation in order to conduct unsupervised paraphrasing. The intention is that this will prevent the model from learning to produce extractive summaries during fine-tuning. We implemented back-translation with the pre-trained OPUS-MT transformer based machine translation model [14]. Specifically, we downloaded weights for the English-to-German, and the German-to-English models.

4.4 Fine-Tuning for Summary Generation

In order to produce summaries using the unsupervised method described above, we leverage the pre-trained BART-base [2] language model. This model performs at or near the state-of-the-art for document summarization, among other language generation tasks. We fine-tune BART using variations of ExtraPhraseRank, then use this fine-tuned model to generate novel abstractive summaries.

Figure 3: The BART Model



The BART model combines a BERT-like encoder with a GPT-like decoder. Image from [2].

4.5 Baselines

As a baseline, we use a heuristic for creating length-controlled abstractive pseudo-summaries, and finetune BART [2] on these summaries. We also compare pseudo-summaries to gold summaries to understand how much room for improvement there is. The way the heuristic generates summaries is as follows: given a desired pseudo-summary length in words, the heuristic returns the first x sentences of the document text, where x is the number of sentences such that the pseudo-summary is closest in length to the desired number of words. We then back-translate the pseudo-summaries as we do in the ExtraPhraseRank algorithm. This allows us to evaluate ExtraPhraseRank, and specifically it's textRank component that differs from ExtraPhrase, in its ability to be used to train BART to produce high quality abstrative summaries.

5 Experiments

5.1 Datasets and Preprocessing

We use two main datasets: the Extreme Summarization (XSum) Dataset [9] and the Amazon Review Dataset [15]. The XSum dataset is a collection of over 200k news articles and summaries of those articles. This dataset was created specifically to evaluate abstractive summarization tasks. We sampled 1k summaries from this datasets and split them into 900 train and 100 test examples. We do the standard summarization task with the XSum dataset. We use the Amazon Review Dataset to form a summarization task by concatenating reviews of the same product in order into a document. We then generate pseudo-summaries for fine-tuning using the data augmentation methods described above. Our final dataset contains 1000 distinct product review-summary pairs.

5.2 Evaluation method

We evaluate the downstream ability of the summary generation technique on the XSum dataset by recording ROUGE-1, -2, and -L scores between predictions and gold summaries. We then qualitatively analyze the fluency and substance of the generated summaries for our modified Amazon Review Dataset.

5.3 Experimental details

We have conducted seven experiments on the XSum summarization task, and have also generated and analyzed summaries for the Amazon Review Dataset. For all of our experiments, we train for 1 epoch using a learning rate of 3e-05 with weight decay 0.1 and batch sizes of 4. Our XSum experiments can be understood as follows.

We first seek to understand the case where we have access to gold summaries. Then, we test strategies for improving accuracy given zero access to gold summaries. This includes directly using a pre-trained BART model, or fine-tuning on synthetic summaries. There are three ways that we create synthetic summaries: applying the heuristic described above (selecting the first few sentences), using the TextRank summarization algorithm alone, or using the ExtraPhraseRank pipeline (TextRank, then back-translation). We experiment fine-tuning BART on all of these types of summaries, and then evaluate against gold summaries.

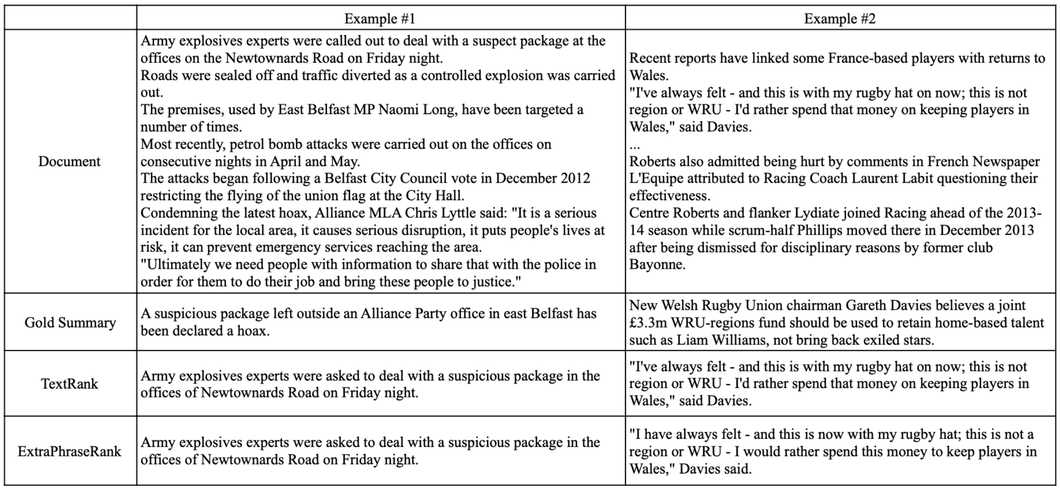
The last thing we seek to determine is whether or not ExtraPhraseRank should be used as a data augmentation strategy in the case that we have *some* gold summaries. To test this, we create a dataset with $1/2$ ExtraPhraseRank-produced synthetic summaries and $1/2$ gold summaries. We then fine-tune BART on this dataset and compare its performance to BART fine-tuned on only these gold summaries.

5.4 Results

5.5 Data Augmentation

ExtraPhraseRank was able to produce pseudo-summaries that resemble human-generated summaries for fine-tuning. Figure 4 depicts the algorithm's outputs when run on examples from the XSum dataset.

Figure 4: ExtraPhraseRank Outputs on XSum Examples



	Example #1	Example #2
Document	Army explosives experts were called out to deal with a suspect package at the offices on the Newtownards Road on Friday night. Roads were sealed off and traffic diverted as a controlled explosion was carried out. The premises, used by East Belfast MP Naomi Long, have been targeted a number of times. Most recently, petrol bomb attacks were carried out on the offices on consecutive nights in April and May. The attacks began following a Belfast City Council vote in December 2012 restricting the flying of the union flag at the City Hall. Condemning the latest hoax, Alliance MLA Chris Lyttle said: "It is a serious incident for the local area, it causes serious disruption, it puts people's lives at risk, it can prevent emergency services reaching the area. "Ultimately we need people with information to share that with the police in order for them to do their job and bring these people to justice."	Recent reports have linked some France-based players with returns to Wales. "I've always felt - and this is with my rugby hat on now; this is not region or WRU - I'd rather spend that money on keeping players in Wales," said Davies. ... Roberts also admitted being hurt by comments in French Newspaper L'Equipe attributed to Racing Coach Laurent Labit questioning their effectiveness. Centre Roberts and flanker Lydiate joined Racing ahead of the 2013-14 season while scrum-half Phillips moved there in December 2013 after being dismissed for disciplinary reasons by former club Bayonne.
Gold Summary	A suspicious package left outside an Alliance Party office in east Belfast has been declared a hoax.	New Welsh Rugby Union chairman Gareth Davies believes a joint £3.3m WRU-regions fund should be used to retain home-based talent such as Liam Williams, not bring back exiled stars.
TextRank	Army explosives experts were asked to deal with a suspicious package in the offices of Newtownards Road on Friday night.	"I've always felt - and this is with my rugby hat on now; this is not region or WRU - I'd rather spend that money on keeping players in Wales," said Davies.
ExtraPhraseRank	Army explosives experts were asked to deal with a suspicious package in the offices of Newtownards Road on Friday night.	"I have always felt - and this is now with my rugby hat; this is not a region or WRU - I would rather spend this money to keep players in Wales," Davies said.

For two different XSum examples, we compare the original document-summary pair to the summary generated by TextRank alone, and then also to the summary generated by ExtraPhraseRank.

5.6 Downstream Results for XSum

Our results show that fine-tuning with ExtraPhraseRank-generated data modestly improves the ability of BART to produce summaries matching human-generated sentences in the case where no training data is available. Still, we find that ExtraPhraseRank performs poorly in the case where some gold summaries exist, decreasing the ROUGE score that the model is able to achieve versus when it is only trained on few gold examples.

Table 1: Experimental Results on XSum Dataset

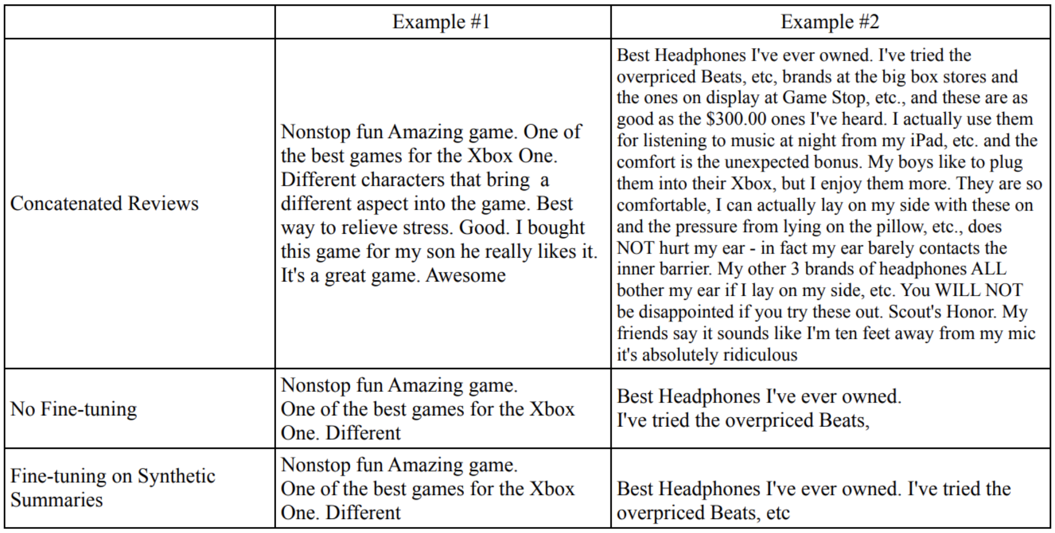
	ROUGE-L	ROUGE-2	ROUGE-1
Gold	**27.39**	**14.021**	**34.091**
No Fine-tuning	10.971	1.311	13.563
Heuristic	10.539	1.046	14.055
TextRank	11.27	1.329	14.374
ExtraPhraseRank	**12.008**	**1.482**	**14.972**
1/2 ExtraPhraseRank & 1/2 Gold	**15.383**	3.32	19.018
1/2 Gold	10.994	**25.202**	**30.768**

This table shows the ROUGE scores produced by a pretrained BART-base model after fine-tuning on a variety of different augmented versions of the XSum dataset described above in the experiments section.

5.7 Downstream Results for Amazon Review Datset

Fine-tuning BART on the augmented Amazon Review Dataset showed little-to-no improvement compared to the baseline BART generation model with no fine-tuning. It appears that the model both before and after fine-tuning only learned to copy the first words of the prompt text. Slight length differences between the generated sentences before and after fine-tuning indicate that the controlled length was possibly learned. The figure below shows two examples of these results.

Figure 5: BART Outputs on Amazon Review Dataset Examples



	Example #1	Example #2
Concatenated Reviews	Nonstop fun Amazing game. One of the best games for the Xbox One. Different characters that bring a different aspect into the game. Best way to relieve stress. Good. I bought this game for my son he really likes it. It's a great game. Awesome	Best Headphones I've ever owned. I've tried the overpriced Beats, etc, brands at the big box stores and the ones on display at Game Stop, etc., and these are as good as the $300.00 ones I've heard. I actually use them for listening to music at night from my iPad, etc. and the comfort is the unexpected bonus. My boys like to plug them into their Xbox, but I enjoy them more. They are so comfortable, I can actually lay on my side with these on and the pressure from lying on the pillow, etc., does NOT hurt my ear - in fact my ear barely contacts the inner barrier. My other 3 brands of headphones ALL bother my ear if I lay on my side, etc. You WILL NOT be disappointed if you try these out. Scout's Honor. My friends say it sounds like I'm ten feet away from my mic it's absolutely ridiculous
No Fine-tuning	Nonstop fun Amazing game. One of the best games for the Xbox One. Different	Best Headphones I've ever owned. I've tried the overpriced Beats,
Fine-tuning on Synthetic Summaries	Nonstop fun Amazing game. One of the best games for the Xbox One. Different	Best Headphones I've ever owned. I've tried the overpriced Beats, etc

For two different Amazon Review Dataset examples, we compare the original document to the summary produced by a vanilla pre-trained BART versus a ExtraPhraseRank fine-tuned BART.

6 Analysis

Fine-tuning BART using variations of the ExtraPhraseRank revealed many things about how well the algorithm worked. Here, we seek to understand the model's performance qualitatively.

- *XSum:* Our experiments using the XSum corpus sought to understand whether ExtraPhraseRank could produce synthetic summaries that could be used to fine-tune a pre-trained BART model to produce better summaries. We found that in a setting where no training data is available, using pseudo-summaries for fine-tuning can improve results. We suspect that this may be due to the length-controlling mechanism of the pseudo-summaries. Concretely, the psuedo-summaries are constructed to be similar in length to the human-generated summaries in xSum. By fine-tuning on these synthetic examples, we coerce model to produce outputs that are similar in length to the reference summaries, thereby increasing ROUGE scores slightly. This is only responsible for part of the gain we see, however, because we also observe that the substance of the ExtraPhraseRank sentences is also better suited for fine-tuning than the Heuristic sentences and than TextRank sentences alone. Therefore, we hypothesize that paraphrasing via back-translation, as opposed to purely extractive data augmentation methods, increases ROUGE scores by inhibiting the model from learning strictly to copy sentences from the source, which in turn drives it to generate sentences more similar to the abstractive sentences found in the XSum dataset.

Moreover, we experimented with using ExtraPhraseRank to supplement, rather than fully replace, existing gold summaries. We found that even using a small number of gold summaries produces significantly higher ROUGE scores than a combination of the two. This is likely because, despite our attempts at paraphrasing, the abstractiveness of XSum summaries is significantly higher than that of ExtraPhraseRank generated pseudo-summaries. Thus, the model performs significantly better when not exposed to these more extractive summaries. A notable exception to this is the ROUGE-L performance of the BART model trained with ExtraPhraseRank summaries. We suspect that this is related to the extractiveness of these summaries, which teaches the model to copy long sections of the input text that may also appear in test summaries.

- *Amazon Review Dataset:* Our experiments using the Amazon Review Dataset studied the practicality of fine-tuning a summarization model based only on a synthetic corpus. In a qualitative analysis of the summaries produced, we found that the BART model largely learned to copy the first few sentences of the source text. The generated summaries to produce have length similar to the average review-length, as was intended by the construction of the corpus. This shows that ExtraPhraseRank is capable of guiding the output length of the summaries without deteriorating their content. However, it is unable to improve the content of summaries directly.

7 Conclusion

Our primary findings are two-fold. First, ExtraPhraseRank represents a viable data augmentation method for controlling the length of summaries generated by large language models. Still, it is limited by its reliance on the extractive TextRank algorithm, and is therefore less capable of replacing human-written gold summaries in the context of fine-tuning language models. We also successfully demonstrate how the length limit problem of unsupervised summarization can be at least addressed via fine-tuning with synthetic data with bounded consequences on sentence quality.

Pre-trained large language models are remarkable at generalizing patterns from even relatively few good training examples. We proved this yet again by demonstrating that, even with half as many examples, well-formed, human-generated summaries outperform synthetic summaries. Among our primary actionable findings is that resources are better spent creating few genuine examples than using data augmentation strategies to create many examples that are mere approximations of the target task.

The biggest limitation of our work is the small amount of training data we were able to use (1000 examples) in comparison to the size of the XSum and Amazon Review Datasets (both over 200,000 examples). This limitation was largely a result of our back-translation algorithm, which relied on very large language models, and was thus computationally expensive to do at scale. This decision was made in order to preserve the integrity of back-translated sentences. Future studies could address this by using more computational resources, or adopting an alternate method of machine translation. Another limitation of our work is our reliance on the XSum dataset. We chose this dataset specifically because of its abstractive nature, yet perhaps a more extractive dataset, such as the popular DailyMail/CNN Summarization Dataset [10], would better showcase the abilities of the ExtraPhraseRank algorithm.

References

[1] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[2] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[3] Mahmood Yousefi-Azar and Len Hamey. Text summarization using unsupervised deep learning. *Expert Systems with Applications*, 68:93–105, 2017.

[4] Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[5] Mengsay Loem, Sho Takase, Masahiro Kaneko, and Naoaki Okazaki. Extraphrase: Efficient data augmentation for abstractive summarization. *arXiv preprint arXiv:2201.05313*, 2022.

[6] Alexander Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[7] Sandeep Subramanian, Raymond Li, Jonathan Pilault, and Christopher Pal. On extractive and abstractive neural document summarization with transformer language models. *arXiv preprint arXiv:1909.03186*, 2020.

[8] Ignacio Tampe, Marcelo Mendoza, and Evangelos Milios. Neural abstractive unsupervised summarization of online news discussions. *arXiv preprint arXiv:2106.03953*, 2021.

[9] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018.

[10] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606, 2016.

[11] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

[12] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[13] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*, 2016.

[14] Tomoki Hayashi, Shinji Watanabe, Yu Zhang, Tomoki Toda, Takaaki Hori, Ramon Astudillo, and Kazuya Takeda. Back-translation-style data augmentation for end-to-end asr. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 426–433. IEEE, 2018.

[15] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, 2019.