APTSumm at BioLaySumm Task 1: Biomedical Breakdown, Improving Readability by Relevancy Based Selection.

A.S. Poornash[†], Atharva Deshmukh[*], Archit Sharma[*], Sriparna Saha
Department of Computer Science & Engineering
Indian Institute of Technology Patna
Patna, Bihar, India-801106
{poornash_2101cs01, atharva_2101cs14, archit_2101ai05,
sriparna}@iitp.ac.in

Abstract

In this paper we tackle a lay summarization task which aims to produce lay-summary of biomedical articles. BioLaySumm in the BioNLP Workshop at ACL 2023 (Goldsack et al., 2023), has presented us with this lay summarization task for biomedical articles. Our proposed models provide a three-step abstractive approach for summarizing biomedical articles. Our methodology involves breaking down the original document into distinct sections, generating candidate summaries for each subsection, then finally re-ranking and selecting the top-performing paragraph for each section. We run ablation studies to establish that each step in our pipeline is critical for improvement in the quality of lay summary. This model achieved the second-highest rank in terms of readability scores (Luo et al., 2022). Our work distinguishes itself from previous studies by not only considering the content of the paper but also its structure, resulting in more coherent and comprehensible lay summaries. We hope that our model for generating lay summaries of biomedical articles will be a useful resource for individuals across various domains, including academia, industry, and healthcare, who require rapid comprehension of key scientific research.

1 Introduction

The sharing of scientific knowledge and engaging with the public is vital for accelerating the acceptance of science in society. The process of generating lay summaries manually can be time-consuming and challenging, especially for authors who may be unfamiliar with communicating their findings to non-specialist audiences. When scientifically backed information is summarized and conveyed in layman's terms, it empowers people to combat the spread of misinformation. This is especially important in the context of the recent COVID-19 pandemic, where understanding medical information is crucial for everyone.

Our project aims to bridge the gap between the increasing availability of health information and the difficulty the public has understanding it. Working towards this direction we took part in the BioLaySumm 2023. We participated in Task 1 in which given an article's abstract and its important sections as input, the goal is to train a model to generate the lay summary. This shared task also emphasized the importance of paragraph readability in the evaluation of lay summaries (Luo et al., 2022). The training and evaluation of our model were carried out on two distinct datasets, namely PLOS and eLife (Goldsack et al., 2022). These datasets consist of biomedical research articles, along with their corresponding technical abstracts and expert-written lay summaries. The metrics used for evaluation are divided in 3 parts, 1) Relevance : This included ROUGE (1, 2, and L) (Lin, 2004) - recall based metrics and BERTScore (Zhang et al., 2019), 2) Readability : Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975) and Dale-Chall Readability Score (DCRS), (Dale and Chall, 1948) 3) Factuality : BARTScore (Koh et al., 2022). In case of Relevance and Factuality higher scores mean good result, but for Readability lower scores are preferred. The checkpoints for BERTScore and BARTScore were provided by the organization. Considering the potential significance of each section, we are using the structure of the article based on the relevancy, then incorporating a re-ranking based approach, (some of which is discussed in the paper SimCLS (Liu and Liu, 2021)) on the sections chosen on the basis of their relevancy to the target summary. Then joining the section-wise summaries generated in a pre-determined manner to generate the final summary. The pre-determined manner is finalized after various experiments and the method we arrived at effectively captures the essential information. Our

[†]Indicates corresponding author (email-aspoornash2355@gmail.com)
[*]These authors contributed equally to this work.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 579–585

system outperforms existing methods in terms of readability, achieving the second-highest score on this metric. Also, our pipeline performs well on the commonly used relevancy metrics (ROUGE scores) for summarization. We attribute this success to leveraging the document structure and careful selection of the most informative parts in each section.

2 Related Works

In this section we will discuss ideas related to Summarization and Structure based Summarization.

2.1 Automatic Text Summarization

Automatic text summarization is the task of generating a shorter version of a given text while retaining its most important information. There are two main types of automatic text summarization: extractive and abstractive summarization.

2.1.1 Extractive Summarization

Extractive summarization is a technique of automatic text summarization that aims to produce a summary of a given text by selecting a subset of its most important and relevant sentences or phrases. Previous work on extractive summarization includes algorithms such as LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004), which use graph-based methods to identify important sentences in a document. These models serve as preliminary, as their performance is not great when compared to current state of art. The next important types of models are transformer based models which generally have quadratic complexity with respect to the sequence length, thus making them prohibitively expensive for large sequences, in our case the whole article. Hence, more research has been done for processing long documents (Bishop et al., 2022, Xiao et al., 2021).

2.1.2 Abstractive Summarization

Abstractive summarization is a type of automatic text summarization that aims to generate a summary that goes beyond simply pruning the unnecessary sentences from the original document. Abstractive summarization involves understanding the meaning and context of the input text, and using that understanding to generate a new summary. This approach is more challenging than extractive summarization because it requires the model to have a deeper understanding of language and the ability to generate human-like language. But due to the breakthrough of recent Transformer based models (Vaswani et al., 2017) like BERT (Devlin et al., 2019), BART (Lewis et al., 2020), and PEGASUS (Zhang et al., 2020) using these models has become key to achieve good performance in abstractive summarization tasks. With more advancements in recent years we have models which incorporate further features like re-ranking (Ravaut et al., 2022) and contrastive learning (Liu et al., 2022). One notable model that we utilized in our research is SimCLS and it has shown promising results in producing effective summaries. More about it is discussed in the methodology section.

Metric	Abstract		Lay Summary	
	PLOS	eLife	PLOS	eLife
FKGL	15.04	15.57	14.76	10.92
DCRS	16.39	17.68	15.90	12.51

Table 1: Readability scores on different metrics (lower the better) (Goldsack et al., 2022)

2.2 Re-Ranking and Relevancy Based Summarization

Traditional abstractive summarization methods have a limitation, the restriction imposed by their token length, typically limited to 1024 tokens. This prevents them to get the context of the whole long document and thus making them less suitable for lengthy documents. Relevancy based summarization, on the other hand, offers a promising solution to this problem by generating summaries that capture the essential information at different levels of detail based on relevancy. Recent papers (Ruan et al., 2022) have analyzed this to be true for long articles and research papers. Also, re-ranking of sentences/summaries has given very good results for both extractive and abstractive (e.g, SimCLS) model. One of the extractive models (Narayan et al., 2018) used a reinforcement learning based approach to rank sentences to produce a summary. But extractive models aren't the best for lay summarization, so we preferred abstractive. Considering this we have utilized only those sections of the given article that contributed to increasing the overall relevancy of the generated summary w.r.t target summary. The basis of this section selection is discussed later.

3 Datasets

In this shared task, we were given access to two datasets, PLOS and eLife, with the objective of

Combination	ROUGE-1	ROUGE-2	ROUGE-L
Scores in BART			
Abstract (Abs)	**41.79**	**12.39**	**35.52**
Introduction (Intro)	40.16	11.83	33.28
Discussion (Disc)	39.35	9.76	31.35
Result1 (Res1)	36.71	8.21	30.49
Result2 (Res2)	34.88	7.72	30.04
Final Pipeline Scores			
*Abs + Intro	46.27	13.66	43.29
Abs + Intro + Disc	44.78	12.59	41.32
Abs + Intro (ROUGE Maximization)	**48.25**	**14.21**	**45.32**

Table 2: Baseline scores, all scores are reported for validation set.

creating a model for lay summarization. The PLOS dataset consists of 24,773 training instances, 1376 validation instances, and 142 test instances, while the eLife dataset consists of 4346 training instances, 241 validation instances, and 142 test instances (Goldsack et al., 2022).

However, despite these similarities, there are notable differences in the structure and content of the lay summaries and articles present in these datasets. Specifically, the eLife dataset has larger article and lay-summary sizes compared to the PLOS dataset as referred in their release paper (Goldsack et al., 2022) . Additionally, the eLife dataset is much more readable/lay than the PLOS dataset, as measured by FKGL and DCRS metrics (as shown in Table 1). Due to these differences, and also because of our experiments, instead of using a single model for both datasets, we decided to use two separate models, one for each dataset.

In addition to the provided datasets, we also utilized the PubMed dataset (Cohan et al., 2018) for training our model. Most pre-trained models available were trained on the CNN/DailyMail dataset (Nallapati et al., 2016). Using these models directly would have resulted in subpar results as the vocabulary and the size of the summary to generate are very different from that needed for our task. Therefore, we fine-tuned our BART model on the PubMed dataset , which had a similar size and vocabulary as that needed for our task.

4 Methodology

Our pipeline has a three-step abstractive approach for summarizing biomedical articles. Our methodology entails breaking down the original document into distinct sections, selecting relevant sections and then generating candidate summaries for each section, and then re-ranking and selecting each section's k-top-performing candidate-summaries. Last two steps are covered with the help of the model SimCLS with the adjustment of the Seq2Seq model used to generate candidate summaries within it. We subsequently join the top summaries of each section with the purpose of finding the best combination. Overall pipeline is shown in Figure 1).

4.1 Seq2Seq model training and Candidate summaries generation

We adopt the BART model as a Seq2Seq model to generate the candidate summaries. We initialize the model weights with a checkpoint pre-trained on the CNN/Daily Mail dataset. Then to get the the model accustomed to the biomedical vocabulary we trained it on the train split of the Pubmed dataset for 2 epochs. This checkpoint is now then fine-tuned on the selected relevant sections to generate the list of candidate summaries for each such section. In our pipeline we generate 16 candidates for each chosen section utilizing beam search sampling strategy. We have chosen to generate 16 candidates for each data point as according to the original SimCLS paper increasing the number of candidate summaries help the re-ranker to learn better underlying features of the dataset.

4.2 Section-wise breakdown along with relevant section selection

As the input context limit of BART and RoBERTa are 1024 and 512 tokens respectively, we partitioned the article with taking into account its general structure, as five sections - Abstract, Introduction, Result1, Result2 and Discussion each containing a maximum of 1024 tokens. The Seq2Seq checkpoint from the previous training is fine-tuned on the train split of each of the five sections separately and then was evaluated on the validation split of each section again separately. This forms the basis of relevance based section selection. We mainly consider the ROUGE-2 score of the generated summary from the fine-tuned BART on that section against the target summary to assess that section's importance in the final summary. From Table 2) we can see that the descending order of section importance w.r.t target summary is Abstract, Introduction, Discussion, Results1 and Results2. One important observation we made while experimenting was that extending the generated summary to a combination of more than three candidate summaries (from any section) leads to a decrease in all the ROUGE metrics. This is due to the fact that all

* Baseline Score.

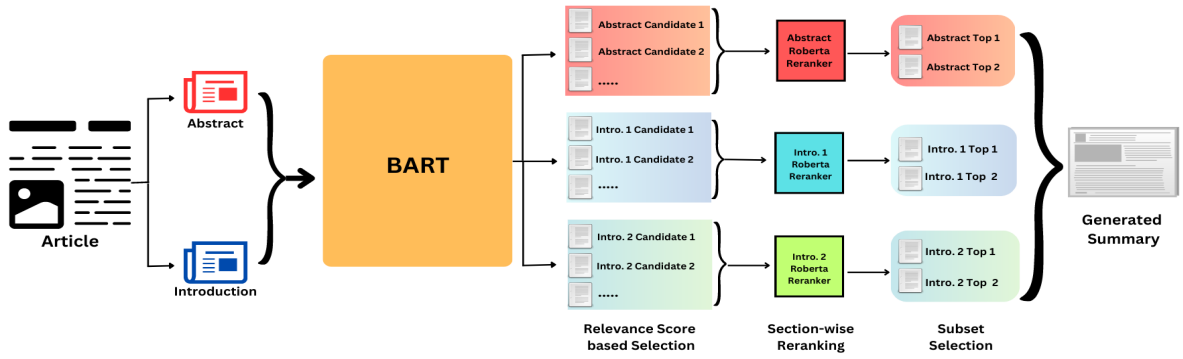

Figure 1: Flow of our complete pipeline. BART is used for Candidate Generation and RoBERTa is used for re-ranking the candidate summaries.

precision of all ROUGE metrics generally tend to decrease when the increase in number of relevant tokens is not able overcome the increase in the total number of tokens of the generated summary. So we tried incorporating the Discussion section also into our pipeline along with Abstract and Introduction Table 2) but that did not perform well in comparison to only using the Abstract and Introduction. Hence we have selected the Abstract and Introduction sections to participate in the generation of the final summary.

4.3 SimCLS and Section Specific Re-rankers

The SimCLS paper entails a two-step process for text summarization. Initially, candidate summaries are generated from a Seq2Seq model. Subsequently, these candidates undergo re-ranking using a custom loss function proposed in the research paper. The loss function is designed to optimize the models performance by creating a ranking list of candidate summaries based on their ROUGE scores in relation to the target summary. This ranking list is then utilized to fine-tune the RoBERTa model during training.

During inference, the fine-tuned RoBERTa model is employed to encode both the source document and the candidate summaries into embeddings. The cosine similarity scores between the embeddings of the source document and each candidate summary are calculated. These scores are used to re-rank the candidates, resulting in a final list of candidates based on their similarity to the source document. Incorporating cosine similarity scores between the RoBERTa embeddings and the source document, potentially enhances the quality of the generated summaries.

In our pipeline we use checkpoints obtained according to Section 4.1's specifications to generate candidate summaries for the relevant sections (Abstract and Introduction). Here we split the 1024 token long Introduction section into two sections to accommodate RoBERTa's input context limit of 512 tokens. This creates two references : Intro 1 and Intro 2 for the same 16 candidates generated from the 1024 token sized Introduction. Then the two copies of Introduction's candidate summaries are re-ranked separately against the two splits of Introduction (Intro1 and Intro2), similar to abstract's candidate summaries. Hence the abstract's candidate summaries are re-ranked against abstract while the two copies of the same introduction are re-ranked against Intro1 and Intro2 as their reference seperately. In the original paper only one RoBERTa based re-ranker was used for the entire document. We adopt a section specific re-ranking strategy that uses a separate re-ranker for each of the final three chosen sections.

4.4 Combination Approach

Now, we have 3 lists of, abstract, Intro 1 and Intro 2 and the 2 top candidates of each section. For all 6 candidates we searched through all 63 ($2^6 - 1$) possible combinations of the six summaries to select the combination, that maximizes the ROUGE scores and improved the readability of the final summary.

We ranked each combination based on ROUGE-1, ROUGE-2, and ROUGE-L scores, getting 3 lists. Final ranking was done by combining the lists where priority was given to ROUGE-2, followed by ROUGE-1, and ROUGE-L. This ranking was done on the validation set where we had access

Combination	Relevance				Readability		Factuality
	Rouge1	Rouge2	RougeL	BERTScore	FKGL	DCRS	BARTScore
Validation Set							
Abstract-Top1 + Abstract-Top2 + Intro1-Top1	**48.25**	**14.25**	**45.24**	84.49	**12.47**	**9.00**	-3.42
Abstract-Top1 + Intro1-Top1 + Intro1-Top2	48.01	14.13	44.89	84.65	12.75	9.05	**-3.41**
Abstract-Top1 + Intro1-Top1 + Intro2-Top1	48.16	14.06	44.95	**84.67**	12.53	9.022	-3.41
Test Set							
Abstract-Top1 + Abstract-Top2 + Intro1-Top1	48.32	14.91	45.41	84.30	12.22	8.99	-3.40

Table 3: Scores of different final combinations.

to the target summaries as well as the generated summaries, in which we obtained Abstract-Top1 + Abstract-Top2 + Intro1-Top1 as our best performing combination across both the datasets. Hence we observed a general trend in the scores of the combinations and submitted the same combination of candidates for the test split as well.

5 Experimental Settings

Fine-tuning of BART ran for 2 epochs per section, the maximum token length was set to 1024, and a batch size of 2 was used. Runtime per epoch of each section in eLife was 6 minutes and for PLOS 35 minutes. Then for candidate generation batch size was set to 8 and, per section it took 13 hours 12 minutes for PLOS and 2 hours 36 minutes for eLife. Finally for re-ranking, the batch size was 8 and per section it took 44 minutes for PLOS and 13 minutes for eLife. For re-ranking portion of inference on the validation and test sets, each sample took 0.11 sec.

6 Results and Discussion

The proposed approach for summarizing biomedical articles in the context of lay summaries has yielded promising results. Here, we present the outcomes of our experiments Table 3) and 2), including the evaluation metrics and analysis of the models performance.

We achieved the second highest rank in terms of readability scores in the BioLaySumm shared task. We hypothesise that this can be attributed to the relevancy based section selection which lead to the selection of Abstract and Introduction sections. Empirically, these two sections are more readable in comparison to the other sections, that go into great domain-specific details, thus generating more readable and lay summaries. We also obtained quite promising relevancy scores on all ROUGE metrics compared to the other submissions of the shared task. However, our pipeline does not perform quite well on the factuality part which is measured by the BARTScore metric. This leads us to conclude that the our generated summaries might be somewhat lacking in factuality. This might be due to the re-ranker giving more priority to relevancy scores than factuality scores, an implementation of the same re-ranking strategy where factuality metric is used instead of the relevance metric might have lead to a improvement in the factuality scores.

The baseline score presented in Table 2 (denoted by an asterisk) were obtained without employing any ROUGE maximization, hyperparameter tuning, or combination selection. These scores were roughly calculated and did not involve optimizing specific hyper-parameters. We also performed ablation studies as shown in Table 2 where, we explored the inclusion of the discussion section in our pipeline but found that it resulted in lower relevancy scores compared to using only the abstract and introduction sections. This suggests that neither the Result1 nor Result2 sections contributed significantly to improving the relevancy scores of the generated summary. In fact, the section-specific scores generated by separate BART checkpoints for Result1 and Result2 were even poorer than those of the discussion section. This implies that the amount of relevant information contained in the candidate summaries of Result1 and Result2 is likely to be less than that found in the discussion section.

7 Conclusion

In conclusion, we proposed a re-ranker based model which leverages the document structure and specifically filters out the most important sections. Our model achieved the second-highest rank in terms of readability scores in the BioLaySumm shared task of the BioNLP Workshop at ACL 2023. We hope that the development of these models will continue to play a critical role in advancing healthcare and will be a valuable resource for individuals across various domains.

Limitations

Our model demonstrates commendable performance in terms of readability and relevancy, but it falls short in the Factuality metric, this is one of the potential areas for improvement. Given more time, one of the directions that we might have explored, is the factuality based re-ranking which considers factuality as metric for comparison, instead of ROUGE scores, or considering both scores giving then certain weights. We have made substantial efforts to improve efficiency and reduce memory requirements, but large language models still impose significant demands on time and computational resources, which remains a limitation of our current work. Additionally, the constraint of a token threshold set at 512 posed challenges in our work. These limitations highlight areas for future research and development.

References

Jennifer Bishop, Qianqian Xie, and Sophia Ananiadou. 2022. GenCompareSum: a hybrid unsupervised summarization method using salience. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 220–240, Dublin, Ireland. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

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. Association for Computational Linguistics.

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.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for*

Computational Linguistics: EMNLP 2022, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, 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. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. 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 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland. Association for Computational Linguistics.

Qian Ruan, Malte Ostendorff, and Georg Rehm. 2022. HiStruct+: Improving extractive text summarization with hierarchical structure information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1292–1308, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. Primer: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv preprint arXiv:2110.08499*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.