Title Generation for Scientific Literature by Self-Attention Sequence Transformer Model

1 Introduction

With the wealth of digital information available today, it is important to be able to quickly and accurately extract important information from large amounts of data [Shi+21]. Within the scientific community, summarization of content into a short title or key tags allows a reader to quickly identify key information that may be relevant to their research interests. In this report, we attempt to title generation from scientific paper abstracts, using variants of the transformer architecture presented in [Vas+17], as a way to test the boundaries of modern abstractive summarization techniques.

Our target dataset is arXiv metadata published by Cornell University, including titles and abstracts for over 1.7 million published research articles [Uni21]. We model this as a sequence-to-sequence (seq2seq) prediction problem, taking tokenized abstracts and predicting tokenized titles. Our results will be evaluated using the BLEU (BiLingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, an automatic method for evaluating machine-generated translations with performance on par with human evaluation [Pap+02; Lin04]. Despite being created for the translation domain, we use these scores due to their prevalence in previous papers on automatic summarization [ML20; Lew+19].

Current work on with text summarization has not gone all the way to title prediction from just the abstract, with the closest known research attempting to predict computer science paper abstracts based on introductions [ML20], or one-sentence summaries of entire papers where the summaries aren't necessarily in the format of a title [Cac+20]. Because examples in our dataset have more compressed information, as the abstract itself is already a compressed version of the rest of the paper, we wanted to expand on previous research to see just how small training examples could get while maintaining good performance.

2 Background

As a baseline approach, we considered a "vanilla" transformer architecture as presented in [Vas+17] and implemented in OpenNMT-py [Kle+17]. After training for just 40000 steps, we achieved 41.98% training accuracy and 42.79% validation accuracy, where "accuracy" represents the proportion of correctly predicted words over the total number of words predicted. A graph of these accuracies over time can be found in Figure 1.

In addition to statistics reported by OpenNMT, we also generated the BLEU and ROUGE scores for each abstract-title pair in the validation dataset, averaging over the whole dataset. We used the python libraries NLTK and rouge for computing the BLEU and ROUGE scores respectively. The baseline achieved a BLEU score of 4.9 and 2-gram F1 ROUGE score of 10.7. This is far below previous results in text summarization using transformers, which have achieved a BLEU score 20.8 and a 2-gram F1 ROUGE score of 41.8 on the CNN/Daily Mail testing dataset [PM19]. While the titles generated were generally coherent, these results suggest that they did not include many of the keywords or key phrases from the reference titles, resulting in low BLEU and ROUGE scores.

Preprint. Under review.

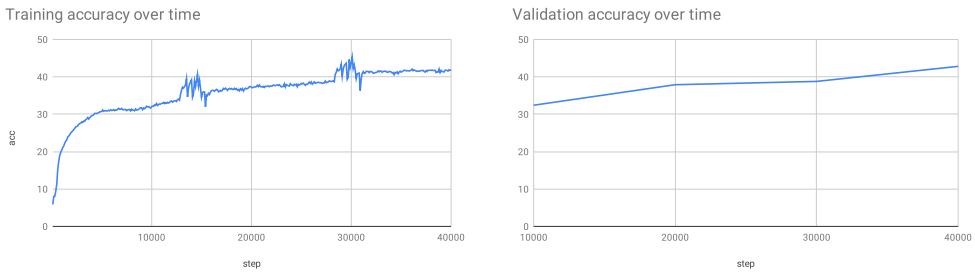

Figure 1: Training and validation accuracy for the baseline method

3 Related Work

3.1 Copy Attention

An early attempt to increase the performance of transformers on summarization tasks was the Pointer-Generator network, a.k.a Copy Attention, presented in [SLM17]. This model included an auxiliary network to predict, given the final decoder hidden state and attention given to the input words, how likely the model is to copy a word from the input. It then used this value to take a weighted average of the input attention distribution and the predicted output distribution to get the distribution to generate the next word [SLM17]. This construction allowed for copying words not in the original decoder vocabulary that nevertheless had high attention values [SLM17].

More recently, a variant called Guided Copy Attention slightly modified how the copy attention distribution was formed, giving rise to even better results [Xu+20]. It used information from the self-attention graph to further boost scores of likely relevant words [Xu+20]. Seeing these good results on summarization using this mechanism guided us to try it ourselves.

3.2 BART

In addition to architectural improvements, other papers showed that training regimes can improve performance on summarization as well. One such paper explored the power of BART, a technique inspired by language modeling that randomly masks/permutes tokens as data augmentation during a pre-training step for seq2seq models [Lew+19]. The authors found that just adding BART to a basic transformer led to moderate increases in BLEU and ROUGE scores for abstractive dialogue, question answering, and and summarization tasks [Lew+19]. The fact that something so simple could lead to such gains was fascinating, so we wanted to try that out at well.

3.3 T5/Transfer learning

Another applicable technique is transfer learning. Transfer learning is when a model is pre-trained on a large, general task before being fine-tuned for a more specific task; this has been shown to achieve state-of-the-art results on text summarization when starting from a model pre-trained on the Colossal Clean Crawled Corpus [Raf+19]. This work was done with the Text-to-Text Transfer Transformer (T5) model [Raf+19], so we also wanted to see how a pre-trained version of it performs on our task.

3.4 Language Models

Lastly, we should also mention work done using language models for abstractive summarization. Language models are pre-trained to minimize the cross-entropy loss of predicting actual pieces of text with an encoder-only structure, and often need fine-tuning to perform well on summarization tasks [LL19]. One approach to fine-tuning language models is to simply concatenate the source and target documents, separating them by a special symbol, and train the model on that new "language". At prediction time, only the source part of the document is provided and the most likely target document as predicted by the model is output. This was the approach taken by GPT-2 and has the benefit of re-using existing structure and training routines, but doesn't get as good of performance as purpose-built models [Rad+19]. Another approach is to randomly initialize a transformer decoder on

top of the pre-trained encoder embeddings, and train the model on the summarization like a seq2seq task. This is the approach taken by BertSumExt and gets good results, but does need to train for longer [LL19]. However, at this time, both approaches have been superseded by BART, so we did not consider them [Lew+19].

4 Methods/Model

Not including the baseline, we tried a total of 4 different models. Our baseline was a standard multi-head self-attention transformer from [Vas+17] as implemented in the OpenNMT-py [Kle+17]. The second model was adding Copy Attention as described in [SLM17], agan implemented in OpenNMT-py. The third was adding BART to the Copy Attention model, using the best set of parameters described: mask ratio of 0.3, using Poisson distribution with $\lambda = 3$ for the lengths, and a 0.1 sentence permutation ratio [Lew+19]. This was also implemented in OpenNMT-py. Finally, we fine-tuned a T5 model from [Raf+19] pre-trained on the Colossal Clean Crawled Corpus (C4) consisting of 750 GB of text data. We tested T5's performance as a summarizer as-is, and after fine-tuning on 48k, 480k, and 960k entries, which accounted for 3.5%, 35% and 70% of the dataset respectively.

4.1 Parameters

The dataset is the same as it was in the midway report and is common among all models. From the data published in [Uni21], we first extracted all the abstract-title pairs, then tokenized them according to a regex that kept all words together, treating punctuation and numbers as separate tokens. Then, the dataset was randomly split into 80/10/10 percent groups for training, validation, and testing data. A random sample of 30000 abstracts from the training set was used to initialize the baseline model's vocabulary. Random samples of 20000 abstract/title pairs were used to initialize the Copy Attention and Copy Attention + BART's shared vocabulary of 50000 tokens.

We used a lot of the same hyperparameters for the baseline, Copy Attention, and Copy Attention + BART models: $N = 6$ transformer layers for the encoder and decoder stacks, with an embedding dimension of $d = 512$, feed-forward network dimension of 2048, and a context of up to 512 tokens. We used the ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.998$, as well as the learning rate schedule described in [Vas+17] with initial $lr = 2.0$ and 8000 warm-up steps. The dropout rate was 0.2 and label smoothing rate was 0.1. For the baseline model, our batch size was 2048 tokens, and for the latter two models we increased that to 4096 tokens due to now having a larger GPU. We chose these hyperparameters based on the ones used in [GDR18], provided with the OpenNMT-py library. Comparison against the validation set was done every 10000 steps. We trained the Copy Attention and Copy Attention + BART models on an AWS g4dn.xlarge Amazon Elastic Compute Cloud (EC2) for 50 hours to 110000 steps and 53 hours to 120000 steps respectively.

For the T5 multi-task model fine-tuning, the same partition of the training, validation and test dataset is used, but instead is tokenized using a pre-trained SentencePiece model [KR18]. We used a vocabulary of 32000 tokens and a maximum sequence length of 512 for downstream training using T5. To fit within our budget and compute memory constraints, we used a version of T5 known as T5 small [Raf+19], training with a batch size of 12 and only doing a partial pass of the dataset (single epoch) limited by our number of steps. As mentioned earlier, we used fine-tuned checkpoints of T5 after training on 0, 48k, 480k, and 960k entries. Unfortunately, we did not train the model for longer due to time constraints.

4.2 Evaluation

After all training was done, we took the appropriate checkpoint for each model and predicted titles for the validation set. The baseline, Copy Attention, and Copy Attention + BART models all used beam search with a beam width of 5 and length penalty 0. For T5, we used a beam width of 2 and length penalty 2.0. Both of these hyperparameters were the defaults for the respective libraries we used to train with. Beam search is used instead of random sampling due to improved performance [SVL14].

Finally, we scored the generated titles against the actual titles using BLEU and ROUGE, two metrics for evaluating similarities between translated or summarized texts and the reference output. BLEU scores are calculated by using the modified unigram precision, where for each word the count of that

word in the prediction and reference are compared, as well as comparing the location of the words in the sentence and the length of the sentence compared to the prediction [Pap+02]. ROUGE scores are calculated by comparing n-grams, sequences of words of length n, that appear in the prediction versus in the reference [Lin04]. ROUGE scores are more recall oriented, while BLEU is a more precision-based measure [Lin04]. The scores for each generated-actual pair were averaged across all pairs scored to get the final values.

We has originally considered running more experiments, including ones that tested Language Models, other combinations of hyperparameters, and a more fine-tuned T5, but could not due to time constraints.

5 Results

Below are the training and validation accuracy and perplexity graphs for the Copy Attention and Copy Attention + BART models we trained. Figure 2 contains the data for Copy Attention, and Figure 3 contains Copy Attention + BART. Note that the training graphs look "spiky" in parts because we ended up not shuffling the data by accident.

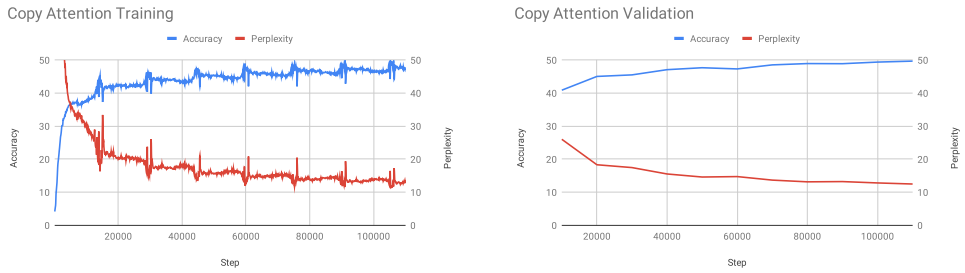

Figure 2: Training and Validation statistics for the Copy Attention method

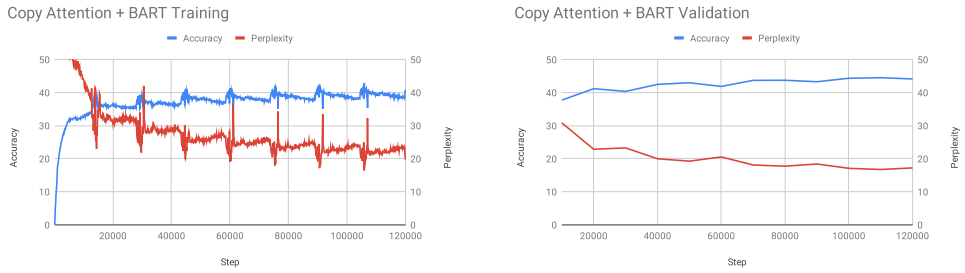

Figure 3: Training and Validation statistics for the Copy Attention + BART methods

For Copy Attention alone, the final OpenNMT accuracy was 47.50% for training and 49.69% for validation, a modest improvement over the baseline method for both partitions. For Copy Attention + BART, the final OpenNMT accuracy was 40.83% for training and 44.15% for validation, a slight decrease from the baseline for the training data but a slight increase from the baseline for the validation data. Looking at figures 4 and 5, there was a large range of outcomes in terms of the quality of the titles generated. Some, such as *Supersymmetric Quantum Mechanics* from Figure 4 were spot on by the model while others, such as *On the index form of a geodesic surface* from Figure 5, versus *Stringy Jacobi fields in Morse theory* were not close at all. Overall, about 1% of titles generated for each model had a ROUGE score greater than 99 (perfect or nearly perfect), which suggests that the model did learn how to predict titles well in specific cases.

As for the BLEU and ROUGE scores, Copy Attention without BART scored 4.8 on BLEU and 16.6 on ROUGE. This BLEU score was almost exactly the same as the baseline, but the ROUGE score increased by over 50%. For the Copy Attention + BART model, the BLEU score increased slightly from the baseline to 5.1, while the ROUGE score only increased to 11.9. The relative comparison of

Reference Title	Copy Attention Generated Title
From dyadic Λ_α to Λ_α	On the Λ_α norm
Supersymmetric Quantum Mechanics	Supersymmetric Quantum Mechanics
Much ado about 248	Three representations of a 248 - dimensional Lie algebra
Soliton surfaces associated with CP^{N-1} sigma models	Soliton surfaces associated with CP^{N-1} sigma models
Polymer Quantum Mechanics and its Continuum Limit	The polymer description of loop quantum gravity

Figure 4: Sample Reference vs Generated Titles - Copy Attention

Reference Title	Copy Attention + BART Generated Title
From dyadic Λ_α to Λ_α	The Λ_α norm of the dyadic grid
Semi - slant submersions	On the semi - slant submersion
Much ado about 248	248 - dimensional Lie algebra of order two
Automatic anomaly detection in high energy collider data	Automatic Anomaly Detection in High Energy Physics
Stringy Jacobi fields in Morse theory	On the index form of a geodesic surface

Figure 5: Sample Reference vs Generated Titles - Copy Attention + BART

scores between models (very similar BLEU scores but much higher ROUGE for the Copy Attention alone model) match up with the discrepancies in Figures 2 and 3, where Copy Attention alone had higher OpenNMT accuracy and lower OpenNMT perplexity than Copy Attention + BART.

For T5, we recorded just the cross-entropy loss shown in Figure 6 with smoothing. Each step in the graph is 500 batches, or 6k abstract-title pairs processed.

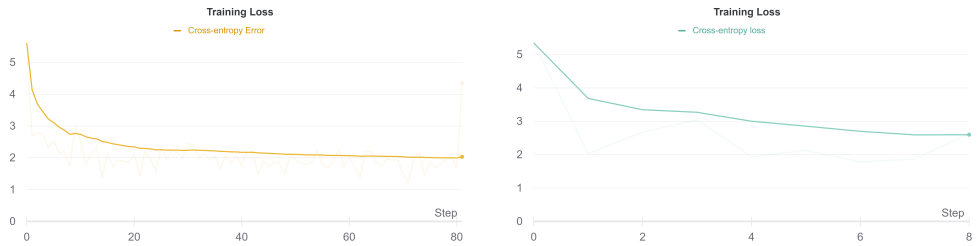

Figure 6: T5 Fine-tune Training progression. (Left) 480k (Right) 48k

Scores \ Size of dataset	non-fine-tuned	3.5%	35%	70%
BLEU	2.5	5.2	4.9	5.1
ROUGE	7.5	14.3	15.9	16.5

Figure 7: Effect of T5 fine-tuning time on performance

For the fine-tuned model, we saw an an impressive gain of ROUGE score of upon fine-tuning with the smallest portion of the dataset. Modest improvement in ROUGE score is observed as the dataset size grows further but not BLEU score as shown in Figure 7. It is possible that it is more difficult to improve abstractive summarization than reducing unnecessary parts in the predictions, as it is observed that the prediction sentence length generally decrease as we increase training. Using only 48k entries or 3.5% of the training set it is able to achieve a BLEU score and ROUGE score of 5.2 and 14.3 respectively, showing just how powerful and generalizable this model is. Sample prediction results from the different T5 checkpoints are shown in Figure 8.

No.		Sentence
1	Ground-truth	Why the laws of physics are just so
	48k	the validity of chemistry in a world that contains chemistry
	480k	physics and general relativity
	960k	the validity of chemistry
2	Ground-truth	Dense Optical Flow based Change Detection Network Robust to Difference of Camera Viewpoints
	48k	Detection of scene changes from a pair of images with a difference of camera viewpoints using dense optical flow
	480k	Detection of Scene Changes from a Couple of Images with a Difference of Camera Points
	960k	Detection of Scene Changes with Dense Optical Flow
3	Ground-truth	Geometry of Fanning Curves in Divisible Grassmannians
	48k	the geometry of fanning curves in $\mathbb{R}$ kn
	480k	the geometry of fanning curves in Grassmann manifolds
	960k	Grassmann manifolds of subspaces of $\mathbb{R}$ kn

Figure 8: Sample Reference vs Generated Titles - T5

6 Discussion and Analysis

6.1 Limitations

The relatively poor results compared to other current research on summarization tasks can likely be attributed to the fact that compression on this scale (abstract to title) has not been attempted before. Compressing an abstract, which is already a compressed form of the entire paper, even further, is likely a very difficult task for any model. Additionally, semantic differences between a title and a general summary can exacerbate this problem. Tools built for summarization and language modeling such as Copy Attention and BART don't necessarily apply to this task; some words in the title appear nowhere in the abstract, reducing the effectiveness of copying, and creating a summary from a summary means that every piece of information in the original summary (the abstract) matters, so techniques which mask information might not be helpful.

Finetuning pre-trained multi-task model on this task is extremely efficient but performance could be improved by optimizing a number of hyperparameters, which we did not do. As shown in Figure 8, the model has a tendency of generating shorter titles the more it was trained. In this instance, we may want to change the length penalty to prevent generated titles from being too short. Changing other parameters such as repetition penalty and the beam width may also help produce better results.

In terms of evaluating the results, while BLEU and ROUGE scores are useful for evaluating the raw similarity of two sentences, they suffer from a couple of drawbacks. For one, it is possible for two sentences to have different words but have the same meaning, for example through substitutions of certain words with synonyms. As these metrics just compare the words in the two sentences [Lin04; Pap+02], this may lead to low BLEU and ROUGE scores if different synonyms are used in the sentences. Additionally, for summarization specifically, the title of a scientific paper isn't necessarily always a direct summarization of the abstract, as it could include concepts not included in the abstract or be modified slightly to try and catch the attention of a reader. One example from the ArXiv dataset is the title *Much ado about 248*, which is about 248-lie Algebra, where the language used in the title is not included in the abstract and highlights the difficulties of summarization problems in general as well as the limitations of BLEU and ROUGE scores. The titles generated by Copy Attention and Copy Attention + Bart, as shown in Figures 4 and 5, seem to accurately represent the topic of the paper in most cases but obviously aren't close to the more creative title decided by the researchers. This highlights the main limitations of BLEU and ROUGE for evaluating translations, since titles generated by the models could be very good titles but would score low in BLEU and ROUGE if they didn't exactly match the vocab and phrasing of the original title.

6.2 Insights

As a whole, this research provides valuable insights into the domain of summarization of scientific papers with the goal of title prediction. As mentioned above, titles are often meant to be creative, and potentially include external information not even included in the abstract. This creates an extremely complex problem that was likely beyond the scope of what the models we used could learn. Our results suggest that Transformer models simply learn language-model-like tasks, and they're not really able to capture external factors such as creativity even when trained on it. Additionally, the structure of titles may be different based off of the field in which the research was published, causing diversity in titles. Incorporating the metadata of the paper containing the year published, author, and field may improve the model performance, but even that may not be enough.

6.3 Future Work

After seeing where the approaches we tried failed, we have a number ideas on how future work could improve these results. First, to address the limitations of BLEU and ROUGE scores, it may be useful to have human scoring of the titles. This would involve crowdsourcing humans reading abstracts and trying to determine which title was generated by a model and which one was the actual title. Humans not being able to tell the two apart would be a good sign that the model had been trained well. This has the benefit of considering context instead of just raw vocabulary, as humans could generally capture the topic of the abstract and tell if it was represented in the title, which a raw comparison of vocabulary cannot.

In terms of the actual modeling, increasing the training time for each model could lead to better results. Also, for the Copy Attention and Copy Attention + BART models, we forgot to shuffle the data, so in future work shuffling the data would probably lead to better results. For T5 specifically, the AWS EC2 instance we used did not have enough memory to train the T5 base model, so we resorted to using T5 small with low batch size. Even then, training time on the entire dataset would have taken way more time than we had, so we had to use only a random subset of the data to train the model. From the results we did get for T5, it seems to show more promise than Copy Attention and Copy Attention + BART. Pre-training probably helped with these results. Overall, we believe that the best area of future work on this topic is in the domain of transfer learning due to the ability of transfer learning models to generalize their learning to downstream tasks in the domain of summarization.

References

[Cac+20] Isabel Cachola et al. "TLDR: Extreme Summarization of Scientific Documents". In: *CoRR* abs/2004.15011 (2020). arXiv: 2004.15011. URL: https://arxiv.org/abs/2004.15011.

[GDR18] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. "Bottom-Up Abstractive Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 4098–4109.

[Kle+17] Guillaume Klein et al. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *Proc. ACL*. 2017. DOI: 10.18653/v1/P17-4012. URL: https://doi.org/10.18653/v1/P17-4012.

[KR18] Taku Kudo and John Richardson. *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. 2018. arXiv: 1808.06226 [cs.CL].

[Lew+19] Mike Lewis et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv: 1910.13461 [cs.CL].

[Lin04] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 2004, pp. 74–81.

[LL19] Yang Liu and Mirella Lapata. *Text Summarization with Pretrained Encoders*. 2019. arXiv: 1908.08345 [cs.CL].

[ML20] Dennis Singh Moirangthem and Minho Lee. "Abstractive summarization of long texts by representing multiple compositionalities with temporal hierarchical pointer generator network". In: *Neural Networks* 124 (2020), pp. 1–11. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2019.12.022. URL: https://www.sciencedirect.com/science/article/pii/S0893608019304228.

[Pap+02] Kishore Papineni et al. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[PM19] Shantipriya Parida and Petr Motlicek. "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)*. 2019, pp. 5994–5998.

[Rad+19] Alec Radford et al. *Language Models are Unsupervised Multitask Learners*. 2019. URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[Raf+19] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *CoRR* abs/1910.10683 (2019). arXiv: 1910.10683. URL: http://arxiv.org/abs/1910.10683.

[Shi+21] Tian Shi et al. "Neural Abstractive Text Summarization with Sequence-to-Sequence Models". In: *ACM/IMS Trans. Data Sci.* 2.1 (Jan. 2021). ISSN: 2691-1922. DOI: 10.1145/3419106. URL: https://doi.org/10.1145/3419106.

[SLM17] Abigail See, Peter J. Liu, and Christopher D. Manning. *Get To The Point: Summarization with Pointer-Generator Networks*. 2017. arXiv: 1704.04368 [cs.CL].

[SVL14] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. *Sequence to Sequence Learning with Neural Networks*. 2014. arXiv: 1409.3215 [cs.CL].

[Uni21] Cornell University. 2021. URL: https://arxiv.org/.

[Vas+17] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[Xu+20] Song Xu et al. "Self-Attention Guided Copy Mechanism for Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 1355–1362. DOI: 10.18653/v1/2020.acl-main.125. URL: https://aclanthology.org/2020.acl-main.125.