Scientific Comparative Argument Generation

Mengxia Yu
University of Notre Dame
myu2@nd.edu

Wenhao Yu
University of Notre Dame
wyu1@nd.edu

Lingbo Tong
University of Notre Dame
ltong2@nd.edu

Meng Jiang
University of Notre Dame
mjiang2@nd.edu

ABSTRACT

In this work, we introduce a new yet important NLP task in scientific domain that is generating comparative arguments that aim to present an invention's technical novelty by comparing it to one or multiple prior works. Any success on this task is a fundamental step towards the goal of enabling machines to think and write like scientists. So we create and release a dataset of good quality and size for benchmarking. We report and analyze the results of advanced text generation models, which uncover the unique challenge of this task compared to traditional argument generation tasks: there is a significant topic gap between inputs and output when the output is comparing instead of summarizing the inputs. We study the impact of the topics on the generation performance and investigate the possibility of learning, predicting, and utilizing the topics. Finally, this work discusses promising directions to achieve the goal.

1 INTRODUCTION

Can machines read, think, and write like scientists? Comparative arguments aim to compare one item against something else and distinguish their similarities and differences, which are widely used to present the technical novelty of the scientist's invention [25]. For the invention to have good scientific or commercial value, it is a significant improvement on prior art [21]. Below is an example: the work of Transformers [20] argued that compared to itself, two papers, factorization tricks [9] and conditional computation [17],

> *"the fundamental constraint of sequential computation, however, remains"*

after summarized them as:

> *"recent work has achieved significant improvements in computational efficiency through factorization tricks [21] and conditional computation [32], while also improving model performance in case of the latter."*

This work presents the first study on automatic scientific comparative argument generation.

We define this problem as follows. Suppose a machine is given a brief description of the idea of an invention, a brief summary of one or multiple prior works, and the information of the prior works. Specifically, the input includes the abstract of paper A (denoted by *Abst_A*), the summary sentence(s) in A citing a set of papers $\mathcal{B}$ (denoted by *Smry*), and the full-text of papers in $\mathcal{B}$ (denoted by $Ftxt\_\mathcal{B}$). The machine aims to generate a comparative argument (denoted by *CmpArg*) that claims the weaknesses of prior works $\mathcal{B}$ and/or technical novelty of A. The above two quoted texts are examples for *CmpArg* and *Smry*, respectively. *CmpArg* often has the word "however".

To benchmark and analyze this problem, we collect 59,765 data examples in computer science from S2ORC [12], employ advanced neural generative models such as BART [10], and evaluate the trained models using both automatic metrics and human. Results reflects good *fluency* and *informativeness* of the generated arguments, however, the models fail to deliver *quality content*: $< .08$ on BLEU-2 score, $< .03$ on BLEU-4, and $< 15\%$ on consistency of the output topics compared to references.

The gap between the topics of input and output texts, in other words, capturing the aspect to compare the multiple input texts, is a unique challenge of comparative argument generation that traditional argument generation tasks do not have [5–8]. For example, counter-arguments or contrastive claims had a different stance or opinion but the same topics as the original argument or claim. In our problem, the common topics between A and $\mathcal{B}$ are the research fields they work on, such as "neural network" and "machine translation"; and the topics of *CmpArg* are the many aspects that prior works have substantial issues or inventions have a substantial improvement. In machine learning, three common aspects are *method assumption*, *labelled/training data amount*, and *computational complexity*.

To further understand the role of topics in comparative argument generation, we recruit machine learning experts to annotate the aspects of comparison. We have 2,639 examples whose *CmpArg* are confidently labelled as one of the three topics. We empirically investigate the possibilities of utilizing and learning the topic information. We observe that if given accurate topics, the topic-guided generation models are able to produce scientific comparative arguments of quality content: $> 76\%$ on topic consistency via human evaluation. Unfortunately, the trained topic classifiers could not predict the topics accurately so the content quality would still be poor when the generation was guided by the predicted topics.

The main contributions of this study are:

- introducing a new yet important argument generation task in scientific domain and releasing a dataset for benchmarking;[1]

KDD '22, August 14-18, 2022, Washington DC, USA

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

[1] Our code and data are publicly available at https://anonymous.4open.science/r/SCAG/

	Average number of words				Data splitting		
	Abst_A	*Ftxt_B*	*Smry*	*CmpArg*	Train	Valid	Test
SCAG	211.7	1709.7	39.2	21.6	53,765	3,000	3,000
SCAGml	200.5	2470.4	48.4	23.1	2,039	300	300

Table 1: Statistics of benchmark datasets.

- uncovering the unique challenge of the task – topic gap and studying its impact on model performance, especially on content quality;
- presenting an opinion with evidence that reasoning the topics of comparative arguments is a nontrivial obstacle of a scientific AI.

2 RELATED WORK

2.1 Scientific Text Generation

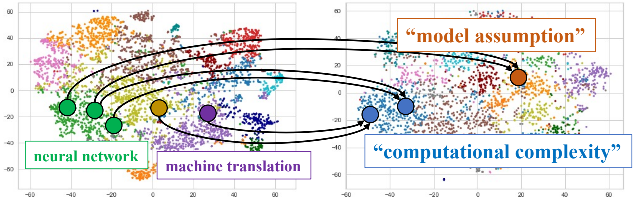
People are curious about whether computers can write scientific papers. SCIgen was a program that generated random computer science papers using a hand-written context-free grammar [15]. PaperRobot [23] was a text-to-text generation system that produced abstract from title, conclusions from the abstract, and next paper’s title from the conclusions. SciGen [13] suggested a table-to-text NLG task and gave a dataset for automatically summarizing experimental results. SciXGen [4] expanded the input to multi-modal objects including tables, figures, theorems, and equations. AutoCite [24] fused multi-modal representations for generating citation sentences. Turing test on PaperRobot showed that a human expert chose the system’s output over human’s output 12–30% time (still much lower than 50%) when they were mixed. The above systems might generate some new statements as fluent sentences which looked different from existing literature and was getting hard to distinguish from human-written texts. However, they are not as intelligent as human research assistants, due to their lack of reasoning abilities about scientific innovation like *making comparisons*.

2.2 Argument Generation

Generating arguments is useful in many domains such as politics, business, and sports. Some approaches considered the task as summarization [22] or making conclusions [18] of argumentative texts. A greater line of work focused on generating arguments with opposite views. To generate them from given claims and stances, Hua et al. combined neural generation model with retriever to create the arguments under the opposite stance. Hidey and McKeown built a sequence-to-sequence model to rewrite claims into contrastive claims. In order to explicitly control the argument content, content planning techniques have been applied to generate counter-arguments [7]. Recent work utilized pre-trained language models to generate arguments conditioned on a specific aspect [16] or a given attribute such as stance and user-belief [1]. Our work is the first attempt to study *comparative* arguments generation, where the arguments are derived by the comparison between two studies.

3 PROBLEM DEFINITION AND BENCHMARKING

A comparative argument is an argument that tries to explain how two subjects are either similar or different. In scientific literature, comparative arguments are often made along with a citation behavior, as compare/contrast is an important citation function [19]. Authors usually cite and summarize the related paper(s) and then make an argument to explain the difference between the cited works and their current work.



citing paper *Abst_A* and cited paper *Ftxt_B* comparative argument *CmpArg*

Figure 1: We use SciBERT embeddings (of 768 dimensions) and t-SNE to represent the input and output texts in two-dimensional spaces, as shown on the left and right, respectively. It’s unlike summarization – comparing two papers of the same or different topics (colored clusters) may lead to different or the same topics in the argument’s embedding space. There is a significant gap between the topics of input and output due to the comparison.

The problem is defined as follows in which the textual variables have been mentioned in Section 1:

Definition (Scientific comparative argument generation). *Given $Abst\_A$, $Ftxt\_\mathcal{B}$, and $Smry$, generate $CmpArg$. The generation model requires to maximize $P(CmpArg|Abst\_A, Ftxt\_\mathcal{B}, Smry)$.*

We collect a benchmark dataset (named SCAG) in a few steps from the S2ORC corpus that contains over 80 million papers from many research fields, with rich information such as paper metadata, abstracts, and citation edges [12]. First, we select papers in Computer Science based on the identifier $field\_of\_study$. Second, we extract pairs of paper and cited paper set $(A, \mathcal{B})$ that satisfy:

- In the full text of A, there is a sentence where A cites a set of papers $\mathcal{B}$ with citation marks.
- Right after *Smry*, there is a sentence has the word “however” that indicates the comparison between A and $\mathcal{B}$.
- The abstract of paper A has more than 50 words.
- The text of each paper $B \in \mathcal{B}$ has more than 50 words.

Then we remove the citation marks from the sentences and replace acronyms of concepts with their complete names. Finally, the SCAG dataset has 59,765 examples. We split it into train, validation, and test sets. Statistics can be found in Table 1.

3.1 Topic Gap Caused by Comparison

Traditionally, generated arguments have the same topics (though of the same or different opinions or stances) with input argumentative text. For example, the summary of a political opinion article should discuss the political topics; the contrast arguments of a restaurant review or a movie review should describe the aspects of restaurants or movies. However, we hypothesize that comparing two documents (especially scientific approaches) may lead to a different topic space of the documents.

Figure 1 visually presents data-driven evidence about our hypothesis. We use a large pre-trained BERT model in scientific domains, called SciBERT [3] to obtain distributed representations of

	Training datasets		Topic guidance	
	SCAG	SCAGml	Predicted	Given
M1	✓			
M2		✓		
M3		✓	✓	
M4		✓		✓
M5	✓	✓		
M6	✓	✓	✓	
M7	✓	✓		✓

Table 2: We develop seven models (M1–M7) based on a pre-trained BART to perform comparative argument generation.

citing papers and cited papers in the SCAG dataset. These representation vectors have 768 dimensions. Then we use t-distributed stochastic neighbor embedding (t-SNE) to project the vectors into a two-dimensional space. We use the K-means clustering algorithm to find the topics in the semantic space. The topics are shown as colored clusters on the left of Figure 1. We do the same operations on the output text, i.e., comparative arguments, which are visualized on the right-hand size of the figure. We observe that comparing two papers of the same or different topics (colored clusters) may lead to different or the same topics in the argument’s embedding space. There is a significant semantic gap between the topics of input and output due to the comparison.

To study the topics of comparison, we recruit *machine learning* experts to annotate the comparative arguments. We require annotators to label their confidence on labelling the topics of arguments and find three common topics that have the highest confidence: (1) method assumption, (2) labelled/training data amount, and (3) computational complexity. Finally, we have 2,039, 300, and 300 annotated examples about *machine learning*. We name this dataset “SCAGml”. Statistics can be found in the second row of Table 1.

4 MODELS

We aim to develop advanced natural language generation (NLG) models to investigate the impact of the data and the labelled output topics on the performance of generating scientific comparative arguments. The settings of seven models are presented in Table 2.

NLG backbone. We choose BART [10] as the backbone model for our study. It is a transformer encoder-decoder model with a bidirectional encoder and an autoregressive decoder. Specifically, we use the pre-trained facebook/bart-base on Hugging Face. Existing studies show that BART is effective when fine-tuned for text generation. The input sequence is the concatenation of *Abst_A*, $Ftxt\_\mathcal{B}$, and *Smry* by a special token “[SEP]”. If there are multiple cited papers ($|\mathcal{B}| > 1$), they are concatenated by a special token “[SEPB]” to make $Ftxt\_\mathcal{B}$. The maximum length of the input sequence is set as 512. We use the complete *Smry* and *Abst_A*, and for each cited paper, we evenly use the frontmost text and cut off the rest. *CmpArg* is supposed to be the output sequence.

Training datasets. SCAGml is a subset of SCAG. In terms of research fields, the articles in SCAGml are about machine learning, and those in SCAG can be broader in computer science. Therefore, when the target test set is in SCAGml, the training set in SCAG can be used as out-of-domain training data for transfer learning. Model M1 is trained on SCAG only. Models M2–M4 use only SCAGml but not SCAG for training. Models M5–M7 are trained first on SCAG and then on SCAGml. By comparing M1 and M2, we will know whether there is a domain gap. By comparing M1 and M5, we will know whether the transfer learning makes a positive impact.

M1 on:	**B-2**	**B-4**	**R-2**	**R-L**	**MTR**
SCAG-valid	8.21	2.90	4.86	20.77	11.48
SCAG-test	7.97	2.90	4.66	20.27	11.36

Table 3: Results of model M1 on SCAG data. Generating comparative arguments is challenging on the large benchmark in computer science due to the topic gap.

Topic-guided design. Suppose we are given the topic of comparative argument when we aim to generate the argument sentence. We put the topic as a special token at the beginning of the input sequence. BART can attend to this token via the encoder’s self attention and decoder’s cross attention. Models such as M3, M4, M6, and M7 adopt the topic-guided design.

Topic prediction. We add a two-layer perceptron to a BART model to predict the topic of arguments. In M3 and M6, the predicted topic is added into the input sequence as a special token. If the accuracy of topic prediction is perfect, the automated generation should have the same quality as the generation that requires a given topic.

5 EXPERIMENTS

We answer three questions in the experiments: Q1) Given *Abst_A*, $Ftxt\_\mathcal{B}$, and *Smry*, can BART generate high quality *CmpArg* when trained and evaluated on the SCAG data? Q2) How much the topic of *CmpArg* can improve the generation of its content? Q3) Can the topic-guided generation be automated by predicting the topics?

5.1 Evaluation Methods

Automatic evaluation metrics. We report BLEU-2 (B-2), BLEU-4 (B-4), ROUGE-2 recall (R-2), ROUGE-L F1 (R-L), and METEOR (MTR) [2, 11, 14]. BLEU is based on n-gram precision. ROUGE is a recall-oriented metric. METEOR measures unigram precision and recall by considering synonyms.

Human evaluation methods. We sampled 20 examples for each comparison aspect (i.e., argument’s topic) from the test set of SCAGml. We recruit 10 human experts in computer science to provide two evaluations for each model generation. The annotators are given the argument reference and asked to review the generation from the following three perspectives: (1) **fluency**–denotes grammatical fluency, (2) **informativeness**–measures the amount of information, and (3) **content quality**–denotes the appropriateness of the comparison aspects and the viewpoints. The aspects refer to the three topics, and the viewpoints refer to the concrete semantic information in the argument. For fluency and informativeness, the annotators are asked to rate the arguments on a Likert scale from 1 (worst) to 3 (best). We provide descriptions and sample arguments for each scale. For content quality, we have the annotators choose one from the following options: (1) “doesn’t have any aspects in Reference”, (2) “have some but not all aspects in Reference”, (3) “have all the aspects but not all the viewpoints in Reference”, and (4) “have all the aspects and viewpoints in Reference”.

	SCAGml-valid					**SCAGml-test**				
	B-2	**B-4**	**R-2**	**R-L**	**MTR**	**B-2**	**B-4**	**R-2**	**R-L**	**MTR**
M1	5.90	2.30	4.00	15.97	11.49	5.36	1.82	3.79	15.54	11.14
M2	15.14	5.96	10.91	27.28	13.01	14.27	5.56	10.66	27.96	12.68
M3	15.28	5.87	11.05	27.48	12.94	13.89	5.45	10.66	27.19	12.29
M4	16.22	6.53	11.62	28.46	13.64	15.66	6.18	11.65	29.18	13.37
M5	16.00	5.87	11.74	28.79	13.54	14.65	6.19	12.15	28.56	12.81
M6	15.89	6.55	11.50	28.46	13.44	14.81	6.11	11.67	28.30	12.77
M7	**16.88**	**6.51**	**12.51**	**30.26**	**14.35**	**16.19**	**6.75**	**12.75**	**30.18**	**13.91**

Table 4: Results of the seven models on SCAGml data. We have three observations: (1) transfer learning is helpful (M5 vs M2, M6 vs M3, M7 vs M4); (2) topic-guided generation with the accurate topic of arguments is effective (M4 vs M2, M7 vs M5); (3) generation guided by predicted topics doesn't perform well (M3 vs M2/M4, M6 vs M5/M7).

	Fluency	**Informativeness**	**Content Quality**
M1	2.86	2.41	14.7%
M2	2.96	2.37	29.2%
M4	2.90	2.44	57.2%
M5	**2.98**	2.37	37.8%
M7	2.96	**2.46**	**76.1%**

Table 5: Human evaluation results.

5.2 Automatic Evaluation Results on SCAG

We investigate whether the language models can generate high quality *CmpArg* after fine-tuned end-to-end on a large amount of data in SCAG. Table 3 shows that the scores are low. For examples, B-2 is lower than 0.08, and R-2 is lower than 0.05. So, unfortunately the answer to **Q1** is "no" – BART cannot generate high-quality comparative arguments when fine-tuned and evaluated on SCAG.

5.3 Automatic Evaluation Results on SCAGml

Table 4 shows the automatic evaluation results.

Transfer learning from out-of-domain data is helpful. We are curious if training with large data with various undefined topics (i.e., SCAG) can help the text generation on a small, topic-constrained dataset (i.e., SCAGml). We adopt continuous training (a basic strategy for transfer learning) in M5, M6, and M7 models. We first train the models on SCAG, fine the best checkpoint, and then continue training them on SCAGml from it. We compare them against M2, M3, M4 models, respectively. Results show that the models that are warmed up by training on SCAG consistently perform better.

Topic-guided generation is effective. In SCAGml every training example has a labelled topic of the output argument. In the models M4 and M7, we use this oracle topic to guide the argument generation. We compare them against the models M2 and M6 that do not have the topic guidance, respectively. The results show that on both the validation and test sets, M4 and M7 outperform M2 and M6, respectively. To answer **Q2**, the oracle topic guidance can bridge topic gap and improve argument generation.

BART and BERT cannot accurately predict the topic of comparative arguments. Given a concatenation of *Abst_A*, *Ftxt_ℬ*, and *Smry* as input, we train sequence classification models based on BART or BERT to classify the comparison aspect into the three labelled categories. Unfortunately, the models achieves a micro-F1 of ∼0.68 on the validation set and a micro-F1 of ∼0.67 on the test set. We are concerned on that they are not good enough to guide the generation.

As shown in Table 4, the models M3 and M6 that are guided by the predicted topics do not perform better or significantly better than (1) M2 and M5 which do not adopt topic guidance at all or (2) M4 and M7 that are guided by the oracle topics. The answer to **Q3** is "probably no" – predicting the topics directly with a sequence classifier is not accurate enough to be an effective topic guidance.

5.4 Human Evaluation Results on SCAGml

Fluency and informativeness. All five models get similar and high scores in terms of fluency. Among them, M5 (transferred from SCAG) gets the highest score of 2.98 out of 3. However, the worst model, M1, still gets a score of 2.86. This shows that the language model BART is capable of generating fluent and grammatically correct texts, so fluency is not an essential issue in our problem. As for informativeness, we observe that the scores of all the five models are also of a small variance.

Content quality. We ask human evaluators to judge whether the generations have the comparison aspects and viewpoints of the references written by the original authors. From Table 5 we observe that only 29.2% of M2's generations have all the comparison aspects of the references. With oracle topic guidance, M4 gets 57.2%, which is improved by +28%. With transfer learning, M5 also outperforms M2 with a score of 37.8%, but the improvement of M5 is relatively smaller than topic guidance. With both topic guidance and transfer learning, M7 gets the highest score of 76.1%, which is improved by +47% when compared to M2. Therefore, we claim that the aspects to make comparison play an important role in generating comparative arguments. The key challenge of comparative argument generation is to accurately identify aspects and viewpoints. Extra knowledge data or advanced models are needed to achieve this goal.

6 CONCLUSIONS AND FUTURE WORK

In this work, we introduced a novel yet important NLP task called scientific comparative argument generation. We created and released a dataset SCAG for benchmarking, and created the SCAGml dataset with labelled output argument topics. On SCAGml, empirical studies revealed that the input-output topic gap is the key challenge of the task by showing: (1) fine-tuning BART cannot directly yield high quality generations; (2) topic guidance significant improves the content quality of argument generation; (3) inferring output topics given input texts is challenging. Future work can explore creative approaches to identify the output topics, such as incorporating citation graphs, integrating knowledge data, and leveraging retrieval augmentation techniques, to look for evidence related to the inputs. Nevertheless, we believe that SCAG will inspire the future work of comparative argument generation, from reasoning, scientific text analysis, to generation of coherent and correct text.

REFERENCES

[1] Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based Generation of Argumentative Claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.* 224–233.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.* 65–72.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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).* 3615–3620.

[4] Hong Chen, Hiroya Takamura, and Hideki Nakayama. 2021. SciXGen: A Scientific Paper Dataset for Context-Aware Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021.* 1483–1492.

[5] Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation.* 54–64.

[6] Christopher Hidey and Kathleen McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. 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).* 1756–1767.

[7] Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument Generation with Retrieval, Planning, and Realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 2661–2672.

[8] Xinyu Hua and Lu Wang. 2018. Neural Argument Generation Augmented with Externally Retrieved Evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 219–230.

[9] Oleksii Kuchaiev and Boris Ginsburg. 2017. Factorization tricks for LSTM networks. *arXiv preprint arXiv:1703.10722* (2017).

[10] 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.* 7871–7880.

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

[12] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Online, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

[13] Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. SciGen: a Dataset for Reasoning-Aware Text Generation from Scientific Tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).*

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics.* 311–318.

[15] Jeffrey M Perkel. 2005. Need a paper? Fake it. *The Scientist* 19, 9 (2005), 12–14.

[16] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-Controlled Neural Argument Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 380–396.

[17] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017).

[18] Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating Informative Conclusions for Argumentative Texts. In *ACL/IJCNLP (Findings).*

[19] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing.* 103–110.

[20] 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 (2017).

[21] Reinhilde Veugelers and Jian Wang. 2019. Scientific novelty and technological impact. *Research Policy* 48, 6 (2019), 1362–1372.

[22] Lu Wang and Wang Ling. 2016. Neural Network-Based Abstract Generation for Opinions and Arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 47–57.

[23] Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2020. Paperrobot: Incremental draft generation of scientific ideas. In *57th Annual Meeting of the Association for Computational Linguistics (ACL).* Association for Computational Linguistics (ACL), 1980–1991.

[24] Qingqin Wang, Yun Xiong, Yao Zhang, Jiawei Zhang, and Yangyong Zhu. 2021. AutoCite: Multi-Modal Representation Fusion for Contextual Citation Generation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining.* 788–796.

[25] Jay Yagnik, Dennis Strelow, David A Ross, and Ruei-sung Lin. 2011. The power of comparative reasoning. In *International Conference on Computer Vision (ICCV).* IEEE, 2431–2438.