Automatic generation of reviews of scientific papers

Anna Nikiforovskaya*†, Nikolai Kapralov‡, Anna Vlasova§, Oleg Shpynov¶‖ and Aleksei Shpilman*¶

*National Research University Higher School of Economics, Saint Petersburg, Russia
†IDMC, Université de Lorraine, Nancy, France
‡Sechenov Institute of Evolutionary Physiology and Biochemistry RAS, Saint Petersburg, Russia
§Saint Petersburg State University, Saint Petersburg, Russia
¶JetBrains Research, Saint Petersburg, Russia
‖Corresponding author, Email: os@jetbrains.com

Abstract—With an ever-increasing number of scientific papers published each year, it becomes more difficult for researchers to explore a field that they are not closely familiar with already. This greatly inhibits the potential for cross-disciplinary research. A traditional introduction into an area may come in the form of a review paper. However, not all areas and sub-areas have a current review. In this paper, we present a method for the automatic generation of a review paper corresponding to a user-defined query. This method consists of two main parts. The first part identifies key papers in the area by their bibliometric parameters, such as a graph of co-citations. The second stage uses a BERT based architecture that we train on existing reviews for extractive summarization of these key papers. We describe the general pipeline of our method and some implementation details and present both automatic and expert evaluations on the PubMed dataset.

Keywords: extractive summarization, natural language processing, BERT language model, scientific papers analysis.

1. Introduction

When approaching a subject that they are not familiar with, a researcher often starts with a search of relevant papers. For example, Google Scholar and Scopus are commonly used tools to search for scientific papers [1]. Besides the search by title, author names, or keywords, these search engines also provide the user with different statistics, like citations of the paper. However, more often than not, it is hard to organize the information from these papers, especially with the current rise in publication numbers. For instance, regarding the latest COVID-19 epidemic, almost 50000 papers were produced in the last six months.

Some automatic tools approach the task from the bibliometric point of view. One example is bibliometric maps that can be built [2]. These maps visualize the way scientific papers are related in the chosen scientific area using extra information about the paper, like its authors, place of publication, and papers it is cited in. Bibliometric methods allow for highlighting the most important or interesting papers in the selected area. Bibliometric methods often use citations as a measure of scientific impact or paper importance in the area. However, it was shown that the author and the place of publication of the paper affect the number of citations [3]. Also, the meaning of citation is actively studied. For example, it has been shown that there are 15 different meanings of the citation [4].

Another aspect of the paper analysis is the summarization of the scientific papers [5]. Studies in this area show that the citation context, i.e., the text surrounding the link to the paper, can be used for its summarization [6]. Moreover, it was demonstrated that citation context reflects the meaning of the paper better than the paper's abstract, and that citation themselves can be used for paper summarization [7]. However, to our knowledge, no studies exist that present an approach that would summarize a group of papers on a specific topic.

When the area of research is established, a researcher might instead refer to a textbook or a review paper. These sources have information already processed and neatly organized. The enormous amount of papers in popular research areas makes it hard to explore the scientific achievements and to discover possible future research directions. Therefore, automatic tools for scientific papers analysis are demanded by researchers and hence are being developed in various areas.

In this paper, we present a novel method for automatic review generation that combines bibliometric analysis of a specific research area to identify key papers together with a BERT-based deep neural network trained to extract the most relevant sentence from these key papers. The result is a tool able to automatically generate a review based on a query from the user.

We also asked experts in various biological areas to evaluate our tool applied to the PubMed database [8], and the results demonstrate that generated reviews are indeed relevant to posted queries.

The rest of the paper is organized as follows. First, we describe relevant work in the areas of bibliometry and automatic summarization. Then we detail our method, along with the preprocessing process we applied to the PubMed database. Lastly, we present the results of automatic and expert evaluation and conclude the paper.

2. Related Work

2.1. Bibliometric analysis of scientific papers

Historically, bibliometrics has arisen from the statistical studies of bibliographies [9]. Nowadays, it can be applied to all sorts of publications, from scientific papers and books [10] to newspapers and patents [11]. Statistics, such as the identification of authors with the highest number of publications and countries with the highest contribution to the research field, can be obtained from such analysis.

Most of the bibliometric methods are based on paper similarity, which can be determined by co-citation analysis, bibliographic coupling, direct citation, and a bibliographic coupling-based citation-text hybrid approach. Moreover, a citation graph can be used to investigate the total citations number of a paper and its dynamics. Algorithms like PageRank [12] are capable of detecting the most exciting or revolutionary articles, that can change the direction of studies in any particular field. Co-citation graphs are used for cluster analysis and identifying dense communities of similar papers. Bibliographic coupling uses a number of common citations as similarity metrics between papers. Hybrid methods imply all of these and have been shown in recent studies to be the most prominent ones since they allow us to detect similar papers and represent the research front in an unbiased manner [13].

Different tools are available for bibliometric analysis: standalone desktop applications like VosViewer [14] or packages for various programming languages such as Bibliometrics package [15] for the R programming language. Another group is websites like Google Scholar offering search services on a particular subject and the ability to perform citation analysis. Dedicated citations indexes databases, like Scopus or Web of Science, can be used to export bibliographic data for a batch of papers. Altogether, existing bibliometric tools either require manual data processing or provide limited analysis capabilities.

The majority of tools outputs paper's abstract to a user. Reading several dozens of abstracts that often describe the results of the paper in broad terms and contain additional, often redundant information may be cumbersome for a user. We believe that sentences from reviews that cite the paper might be a better alternative. However, since not all papers have associated reviews, we need a way for the automatic generation of review-like sentences.

2.2. Extractive summarization

Text summarization approaches can be abstractive or extractive. Abstractive summarization attempts to produce summaries close to the ones a human expert would make. Those summaries might contain new phrases and sentences, which do not occur in the original text. Extractive summarization produces the text composed from sentences of the original text. In this study, we use extractive summarization, as it guarantees that the generation process would not corrupt the information. In this way, we can find the exact sentence in the paper if we need to see its context

Most extractive summarization methods can be presented as a three-step process [5]. First, they create an intermediate representation of the sentences, then they score sentences based on that representation, and finally generate the summary based on those scores.

Lately, deep learning methods are used more and more, including in the task of extractive summarization [16], [17]. These methods use recurrent neural networks or convolutional neural networks to evaluate sentences. In the process of training, these networks create vector representations of sentences. It is also important to note that vector representations can be acquired by training the network to solve an auxiliary task, such as creating a language model [18]. This helps in the case of limited data availability.

In 2018, a new deep neural network-based language model named BERT was introduced [19]. Using this model for natural language processing tasks has led to significant improvements in the results. In particular, in 2019, it was shown that BERT modification for extractive summarization (BERTSUM) is superior in quality to standard machine learning methods and previously available deep learning methods [20]. We have chosen this method as our base model and describe it and our modifications in section 3.3. Figure 3 shows the architecture of this method.
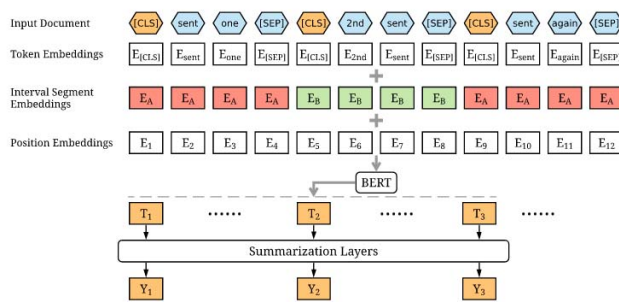

Figure 1. BERTSUM deep summarization network (from [20]).

Early experiments in the summarization of scientific papers mainly concentrate on abstract generation [21] by using the structure of the article and the general rules for dividing papers into sections.

However, it was later shown that *citation contexts* of the paper better reflect the paper content, especially if the information from them is selected properly [22]. *Citation contexts* are sentences from other papers that describe the contents of the target paper. These sentences can be easily identified by the presence of the link to the target paper.

Most works in summarization evaluate the generated summary by comparing it to the reference summary, written by the expert (i.e., the paper abstract). Two commonly used metrics to compare the generated summary and the reference summary are ROUGE and BLEU. They are both based on n-grams count. However, it was shown that in summarization

315

tasks, ROUGE better correlates with human evaluation than BLEU [23]. ROUGE metric is defined as follows:

$$\text{ROUGE-n} = \frac{\displaystyle\sum_{S \in \{\text{reference texts}\}} \sum_{gram_i \in S} Count_{match}(gram_i)}{\displaystyle\sum_{S \in \{\text{reference texts}\}} \sum_{gram_i \in S} Count(gram_i)},$$

where $Count_{match}(gram_i)$ is the number of times n-gram $gram_i$ occurs both in the reference text S and in the generated text, while $Count(gram_i)$ is the number of times n-gram $gram_i$ occurs in reference text S. In our study we use $\text{ROUGE} = \frac{1}{2}(\text{ROUGE-1} + \text{ROUGE-2})$ to compare generated summaries and reference ones.

Automatic evaluation techniques rarely show human satisfaction with the generated summaries. To this end, human evaluation is often employed for a better quality control [5], [24]. In our work, we use both automatic and human expert evaluation of the generated summary.

3. Methods

Figure 2 shows the general pipeline of our automatic review generation method. In this section, we will describe every stage of that process in detail, but we start from the description of the data used.

3.1. Data preprocessing

Our method can be used together with any sufficiently large corpora of scientific papers collected for research in bibliometry and related areas [5], [25].

For this work, we chose PubMedCentral Author Manuscript Collection [8] that contains over 600 000 papers, mostly in areas of biology and medicine. Articles are stored in Journal Article Tag Suite XML format [26]. The data entry includes the paper's unique identifier (PMID) title, abstract, main text, table and figures captions, authors, PMIDs of cited papers, and publication. All this additional information can help in this study, which makes this corpus the most convenient.

Each paper in the corpus has a type. It can be either a general research paper or a review paper, which summarizes some specific scientific areas. There are 8000 review papers out of 600 000 total papers.

We preprocess the raw data into the following data tables. Each has a PMID (a unique identifier of the paper) as key:

- Lists of sentences of each paper that we can use to quickly find all sentences of a paper or sentences under certain numbers. It contains sentence's number in the paper, and the text of the sentence itself.
- Abstracts.
- Image caption texts.
- Table caption texts.
- List of review paper PMIDs. Lists PMIDs of all review papers that cite this paper.

- Citations. PMIDs of cited papers.

An additional table with "reverse citations", that allows for a quick search of citation context by cited paper. With these tables, we can easily access a paper and all associated papers by merging related tables by a specific PMID.

3.2. Citation graph analysis

We developed a tool that combines the capabilities of paper search engines with bibliometrics analysis. The service crawls the PubMed and Semantic Scholar databases as bibliographic data providers and keeps up-to-date information in the Neo4j graph database with daily updates.

The automatic review pipeline starts with a search query by keywords or phrases, where users can choose the most recent, the most cited, or the most relevant papers and limit the size of search results. A citation graph is built on-the-fly and shows overall publication dynamics in time and is used to detect the most popular articles, authors, and journals. We use a hybrid approach for papers similarity determination - a combination of citation, co-citation graph features, bibliographic coupling, and text-based similarity based on the TF-IDF metric [27]. This approach yields excellent results in the case of well-established research areas where citations and co-citations graphs contain a lot of information, as well as in arising topics such as Covid-19, where citations graph is almost empty due to the recency of the papers.

We employ one of the most popular algorithms for community detection - Louvain [28] algorithm to extract closely related groups of papers (topics) from the similarity graph. For each topic, the service shows a word cloud of topic-specific keywords and detailed information about included papers.

Analysis results are combined and presented as an automatically generated review compiled from most review-like sentences from target papers. This review provides a bird-eye view of the research area. It is presented as a table with sentences, quality scores, and additional information about original articles, including topic, publication year, citation number, and a digital online identifier.

3.3. Review-based extractive summarization

As previously mentioned, we base our extractive summarization method on BERTSUM introduced by Liu Yang [20]. The model introduced takes in several sentences separated by specific tags "[CLS]" and "[SEP]" that mark the beginning and end of the fragments of interest, i.e., the sentences. For each input sentence i the BERTSUM model produces a vector T_i, which will be further perceived as a vector representation of the original sentence. Then these vectors are transferred to the linear layer, which will give an estimate of the quality of the sentence for the summary Y_i.

The original model is then trained with binary cross-entropy, as the whole problem was perceived as a binary classification problem, and the goal was to get binary values
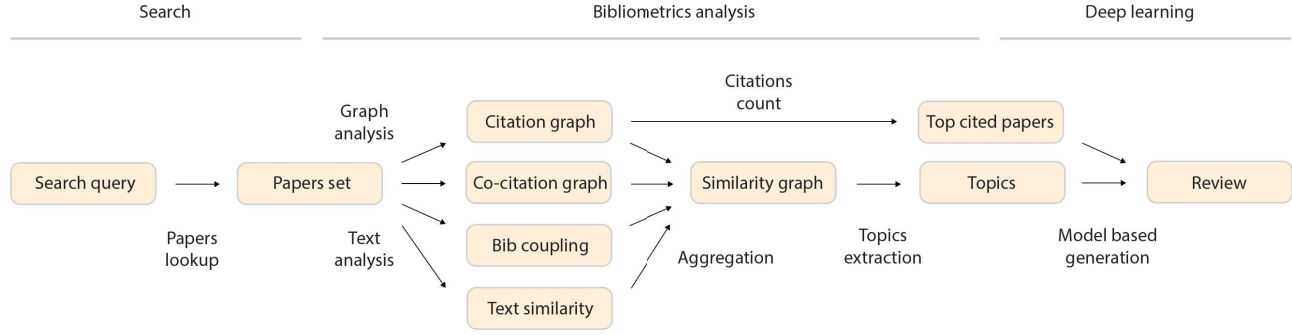
Figure 2. Overall pipeline of our automatic review generation method.

Y_i. In our work, we state the problem differently, as we do not have an ideal set of sentences. We perceive a problem as a regression problem by trying to predict the review-like quality of the sentence, and therefore we have changed the way of training the model.

This way, we consider review papers as "ideal" summaries. However, review papers do not contain the sentences of the cited papers. We can then compare the sentences from the paper with its citation context in review papers to decide if it should be in the summary. We calculate the ROUGE score R_i between the sentence i of the paper and its citation context in review papers. We then use this score as a target value for training the model.

The new loss function during model training is then defined as follows:

$$\text{MSE}(Y, R) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - R_i)^2 \qquad (1)$$

A scientific paper's text is generally much longer than BERT can process during one iteration. BERT model can accept 10-15 sentences per iteration. Due to this fact, we have changed the way of training the model. The text is split into intersecting blocks with the length of the intersection set to 5 sentences.

4. Results

In this section, we present the automatic evaluation of our modified BERTSUM summarization method by examining ROUGE scores of extracted sentences compared to sentences in review papers. We also show the example of a pipeline's output and perform human evaluation by asking experts in specific research areas to grade the output of our method.

4.1. Automatic evaluation

We base the automatic evaluation on the way we state the summarization problem. Since we postulated review papers as "perfect" summaries, we created a test datasets from 300 review papers for benchmarking the summarization.
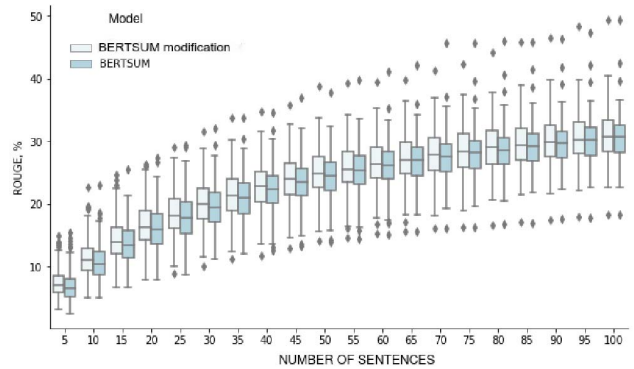

Figure 3. Comparison of non-modified and modified BERTSUM. Our modifications are described in section 3.3.

The final summary is selected by choosing n sentences, given the model's highest scores. Those result summaries are compared to the original review paper via the ROUGE score. The higher the number of sentences, the easier it is to summarize the paper. Hence the ROUGE score generally increases with the increase of n.

We have compared the original BERTSUM and our modified version and present the results in figure 3. Our modified method shows improvement over base BERTSUM in terms of ROUGE score, especially in cases of low n. Since our goal is to create a concise review down to one sentence, if possible, results demonstrate that our modifications help improve the pipeline's overall performance.

4.2. Example output of the pipeline

Table 1 shows an example of output sentences associated with the query "Alzheimer's disease". For every paper, we present the best sentence in terms of the model score is. This score represents how close the sentence to the possible citation context of the paper, i.e., the sentence from a hypothetical review.

As can be seen from the table, our method outputs a diverse collection of papers, covering different aspects of the area of interest. Summary sentences also reflect essential

317

points presented in their corresponding papers, thus helping a potential user to evaluate the state of the research at a glance.

4.3. Expert evaluation

To assess the quality of the result summaries in terms of relevance to the topic of the query and usefulness to the researcher, we have conducted a human expert evaluation on five queries given by experts in specific fields. For each query, the most important 20 papers were selected based on their citations. Then, from each article, one sentence was selected using our modified BERTSUM extractive summarization model.

We asked the experts to classify the generated summary sentences into the three following groups:

- **Not relevant** – the sentence is not relevant to the query.

- **Relevant** – the sentence is relevant to the query, but does not provide useful information about the area.

- **Useful** – the suggestion is useful for understanding the scientific area.

We present the result of this evaluation in table 2. As you can see, the fraction of irrelevant sentences generated is manageably small an 7%, while the average rate of relevant and useful sentences are 33% and 60% respectively. This demonstrates that our method of automatic review generation with extractive summarization produces a diverse review with sentences that can give insight into the queried area of research.

5. Conclusion

In this work, we present a method of automatic review generation by combining bibliometric analysis for key papers identification and deep learning natural language processing by a BERT-based network that evaluates how review-worthy are sentences in identified key papers. The resulting tool generates a list of sentences, each one best describing the result of key papers in response to a query from the user. We evaluate our tool automatically on the PubMed dataset by ROUGE score and manually by asking experts in the field to evaluate specific queries. Both evaluations show that our method can produce relevant one-sentence descriptions of papers.

This tool could be employed to significantly increase researchers' ability to process information in a novel area. It may also assist in writing a traditional review paper or a textbook chapter on a subject.

The code for our method is available at https://github.com/JetBrains-Research/pubtrends-review.

References

[1] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, "Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses," *The FASEB Journal*, vol. 22, no. 2, pp. 338–342, 2008.

[2] M. Cobo, A. López-Herrera, E. Herrera-Viedma, and F. Herrera, "Science mapping software tools: Review, analysis, and cooperative study among tools," *Journal of the American Society for Information Science and Technology*, 2011.

[3] Y. Dong, R. A. Johnson, and N. V. Chawla, "Can scientific impact be predicted?" *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 18–30, 2016.

[4] D. W. Aksnes, L. Langfeldt, and P. Wouters, "Citations, citation indicators, and research quality: An overview of basic concepts and theories," *Sage Open*, vol. 9, no. 1, p. 2158244019829575, 2019.

[5] A. Nenkova and K. McKeown, "Automatic summarization," *Foundations and Trends in Information Retrieval*, vol. 5, pp. 103–233, 2011.

[6] S. Mohammad, B. Dorr, and M. E. et al., "Using citations to generate surveys of scientific paradigms," in *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, 2009, pp. 584–592.

[7] V. Qazvinian, D. Radev, and A. Özgür, "Citation summarization through keyphrase extraction," in *Proceedings of the 23rd international conference on computational linguistics (COLING 2010)*, 2010, pp. 895–903.

[8] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/about/mscollection/

[9] L. Egghe and R. Rousseau, *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Elsevier Science Publishers, 1990.

[10] N. De Bellis, *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. scarecrow press, 2009.

[11] F. Madani and C. Weber, "The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis," *World Patent Information*, vol. 46, pp. 32–48, 2016.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[13] K. W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?" *Journal of the American Society for information Science and Technology*, vol. 61, no. 12, pp. 2389–2404, 2010.

[14] N. J. Van Eck and L. Waltman, "Text mining and visualization using vosviewer," *arXiv preprint arXiv:1109.2058*, 2011.

[15] M. Aria and C. Cuccurullo, "bibliometrix: An r-tool for comprehensive science mapping analysis," *Journal of informetrics*, vol. 11, no. 4, pp. 959–975, 2017.

[16] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," *arXiv preprint arXiv:1603.07252*, 2016.

[17] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[18] Y. Li and T. Yang, "Word embedding for understanding natural language: a survey," in *Guide to Big Data Applications*. Springer, 2018, pp. 83–104.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Paper	Year	Summary sentence	Score
Epidemiology of Alzheimer disease.	2012	The global prevalence of dementia has been estimated to be as high as 24 million, and is predicted to double every 20 years until at least 2040.	0.108
Neuropathological alterations in Alzheimer disease.	2011	Importantly, these cross-sectional neuropathological data have been largely validated by longitudinal in vivo studies using modern imaging biomarkers such as amyloid PET and volumetric MRI.	0.093
Cognitive reserve and Alzheimer disease.	2006	There is also the possibility that directly enhancing CR may help forestall the diagnosis of AD.	0.087
Oxidative stress and Alzheimer disease.	2000	Research in the field of molecular biology has helped to provide a better understanding of both the cascade of biochemical events that occurs with Alzheimer disease (AD) and the heterogeneous nature of the disease.	0.091
Systemic inflammation and disease progression in Alzheimer disease.	2009	OBJECTIVE To determine whether acute episodes of systemic inflammation associated with increased TNF-alpha would be associated with long-term cognitive decline in a prospective cohort study of subjects with Alzheimer disease.	0.115
Classification and basic pathology of Alzheimer disease	2009	Extracellular $A\beta$ accumulation occurs in the parenchyma as diffuse, focal or stellate deposits.	0.085
Loneliness and risk of Alzheimer disease.	2007	CONCLUSION Loneliness is associated with an increased risk of late-life dementia but not with its leading causes.	0.071
The Genetics of Alzheimer Disease: Back to the Future	2010	Three decades of genetic research in Alzheimer disease (AD) have substantially broadened our understanding of the pathogenetic mechanisms leading to neurodegeneration and dementia.	0.099
Increased risk of type 2 diabetes in Alzheimer disease.	2004	Both type 2 diabetes (35% vs. 18%; P < 0.05) and IFG (46% vs. 24%; P < 0.01) were more prevalent in Alzheimer disease versus non-Alzheimer disease control subjects, so 81% of cases of Alzheimer disease had either type 2 diabetes or IFG.	0.107
Hyperinsulinemia and risk of Alzheimer disease.	2004	The HR of AD for the highest quartile of insulin compared to the lowest was 1.7 (95% CI: 1.0, 2.7; p for trend = 0.009).	0.073

TABLE 1. EXAMPLE OUTPUT FOR QUERY "ALZHEIMER'S DISEASE".

Query	Not Relevant	Relevant	Useful
klebsiella pneumoniae plasmid virulence	5%	55%	40%
neisseria meningitidis serogroup	10%	35%	55%
psi prion propagation in yeast	10%	20%	70%
brain computer interface	5%	25%	70%
brain stimulation	5%	30%	65%
Mean Values:	7%	33%	60%

TABLE 2. RESULTS OF THE HUMAN EXPERT EVALUATION.

[20] Y. Liu, "Fine-tune bert for extractive summarization," *arXiv preprint arXiv:1903.10318*, 2019.

[21] C. D. Paice and P. A. Jones, "The identification of important concepts in highly structured technical papers," in *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 1993, pp. 69–78.

[22] A. Abu-Jbara and D. Radev, "Coherent citation-based summarization of scientific papers," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 500–509.

[23] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, pp. 150–157.

[24] W. Xiong and D. Litman, "Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews," in *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers*, 2014, pp. 1985–1995.

[25] L. Hunter and K. B. Cohen, "Biomedical language processing: what's beyond pubmed?" *Molecular cell*, vol. 21, no. 5, pp. 589–594, 2006.

[26] [Online]. Available: https://groups.niso.org/apps/group_public/download.php/21030/ANSI-NISO-Z39.96-2019.pdf

[27] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242. New Jersey, USA, 2003, pp. 133–142.

[28] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E*, vol. 76, no. 3, p. 036106, 2007.