Combining Templates and Language Models for the Automatic Creation of Scientific Overviews

Sarah Frank
CERN, Switzerland
Institute of Information Systems and Data Science
Graz University of Technology, Austria
e-mail: sarah.frank@cern.ch

Andreas Wagner
CERN, Switzerland

Christian Guetl
Institute of Information Systems and Data Science
Graz University of Technology, Austria

***Abstract*—The number of scientific publications is increasing at a rate that makes them progressively more impossible to keep up with. Consequently, automatic creation of summaries from a collection of articles could significantly speed up the selection of publications of interest. This paper focuses on the use case of ultra-short summaries to be used for the creation of topic overviews, as often found in journal editorials. We used a combination of a pre-trained language model and templates to create a coherent text summarizing the papers contained within single journal issues. Following this, we conducted two user studies. The results were generally promising, with users preferring the automatically created summary in a majority of cases. Evaluations of the accuracy, coverage, fluency, and informativeness of the summaries showed that most users found them to be good. However, the variation in the evaluation scores was significant both by user and summary. Text quality was shown to be graded differently according to the user's requirements and familiarity with the typical form of this kind of summary. Furthermore, the importance of high-quality base summaries from the language model, as well as a high number of available templates, cannot be overstated.**

Keywords-automatic summarization; hybrid summarization; language models; natural language processing; templates.

I. Introduction

As new information is generated at an increasingly fast pace, automatic summarization has the potential to support a variety of daily tasks. Despite rapid improvements in the field of NLP (Natural Language Processing), including the creation of large language models, common issues with the generated texts remain. Hallucination, that is, the generation of information that is not supported by the input text, can lead to results that misrepresent statements or are completely false in relation to the input data. Furthermore, the lack of explainability of many existing models leads to difficulties when trying to trace a piece of information back to its source [1]. The particular importance of the information to remain consistent with the source text in scientific environments suggests that current transformer-based solutions often do not meet usability requirements [2].

Before transformers were introduced in 2017, automatic summarisation relied on a variety of different methods such as statistical measures [3], graph-based methods [4][5], and templates [6]–[9]. Template-based summarisation methods, in particular, provide a structured framework for text generation that can enforce certain sentence structures, incorporate domain-specific knowledge, or fulfill given form requirements. Although template-based approaches on their own have disadvantages such as lack of flexibility, previous research has combined them with other methods such as transformers for named entity recognition [10] and general encoder/decoder architecture for electronic direct mail subject generation [11], fine-tuning language models [12], and template-aware summary rewriting [13].

Despite a variety of domains that have utilized templates for improved results, automatic summarization of scientific articles has so far not placed a focus on the approach. With new research being published at a rapid pace, tools that summarize a collection of papers may save a considerable amount of time. The combination of templates with transformer models has the potential to create well-formulated summaries that follow a given structure. This makes it an ideal approach for creating overviews of scientific papers, where a consistent layout is often present.

With the aim to use language models in combination with template-based summarization to create scientific structured text, this research aims to create well-formulated summaries that follow a given structure. Summaries of this kind could then be used to create an overview text of multiple scientific papers. The possibility of receiving regular summaries of recently published papers in a particular field of interest would allow researchers to stay up-to-date on current findings without actively having to search for information. In particular, our goal is to create summaries that can be utilized to give users a brief idea of the topic of a paper for use in editorials, on websites, or in newsletters.

In this paper, we evaluate the effect of combining transformer-model-created summaries with templates. The idea is to use this approach to automatically create overviews of multiple papers, including titles, author information, and short summaries within one sentence each. Texts such as these can, for example, find application as editorial summaries, which are often found in special issues of journals. Due to the lack of a suitable dataset for testing and evaluation of the resulting method, we created a test dataset consisting of 13 special issues comprised of 69 papers, collectively. We then evaluated a selection of language models for their single-sentence summaries using these scientific articles. With this approach, it is possible to retain source knowledge for the information given in the summaries, which is particularly important in scientific

 ISBN: 978-1-68558-192-3

environments. Furthermore, the use of templates allows for the adaptation of the summary structure to the specific use case. Finally, this approach simplifies the evaluation of the factual accuracy of the summary in relation to the source text, as the reference document for each short summary is known.

To this end, our aim is to answer the following research questions (RQs):

- **RQ1:** How do existing language models perform when evaluated for the creation of ultra-short summaries?
- **RQ2:** How can templates tailor results for formulaic texts, such as journal editorials, when used in combination with transformer models?
- **RQ3:** How do the resulting summaries perform when evaluated for language and content quality by automatic and manual means?

In Section 2, we will first give an overview of the background of this work, such as automatic summarization and its importance in general, relevant datasets, and evaluation metrics, followed by an elaboration of works utilizing a combination of transformer- and template-based methods. This is followed by Section 3, which gives an explanation of the general approach and development stages, as well as details the implementation. Subsequently, Section 4 presents the results from both the automatic and manual evaluation methods. Finally, Section 5 discusses the findings and their meaning, and Section 6 finishes the paper by detailing possible limitations and future work.

II. Background and Related Work

The development of automatic summarization techniques has gained significant attention due to its wide range of potential applications. This is due in large part to the creation of transformer-based models [14] and the subsequent popularization of LLMs (Large Language Models), of which GPT (Generative Pre-trained Transformer) [15] and BERT (Bidirectional Encoder Representations from Transformers) [16] are arguably among the most well known. The disadvantage of these approaches is that the results are prone to hallucination, which is the generation of information that is not supported by the source material [17]. Even state-of-the-art models had hallucination-based errors in up to 25% of their summaries when [18] evaluated their correctness in 2019. Despite the fact that numerous attempts have been made to solve this problem since [17], hallucination remains a common issue. Ensuring the production of accurate and coherent summaries that capture the essential meaning of the source text remains a complex task [19]. Transformer-based approaches in particular face challenges related to explainability [20], ambiguity [21][22], redundancy [23][24], and avoiding biases [25][26].

Template-based summarization uses predefined patterns or templates for the generation process. These templates specify how the information from the source text should be organized in the summary and can be designed to capture specific types of information, such as key facts, main ideas, supporting evidence, or other relevant elements depending on the domain. One of the main advantages of this approach is its transparency; the use of fixed templates provides an explicit framework for summarization, which allows the resulting summary to remain explainable [27]. Template-based summarization, which is inherently rigid in its utilisation, can be particularly useful in domains where the structure of information is consistent across documents. Ambiguity, variations in writing styles, and changes in document structure can pose challenges. Furthermore, although the potential for domain-specific customization allows the design of templates that align with the specific needs of a particular use case, this need for domain-specific templates also has a limiting effect [28].

Due to their various advantages and disadvantages, NLP research regularly combines different methods to optimize results. In recent years, there has been an increasing interest in combining template-based summarization, in particular, with other techniques such as pre-trained language models and sentiment analysis [12][11][29][30]. Due to this, templates have also been combined with these methods more frequently, with some approaches specifically making use of pre-trained models such as BERT, and others adding additional pre-training or different attention mechanism. With an aim similar to that of this work in a different domain, Bilal et al. used the combination of templates, sentiment analysis, and abstractive summarization to summarize the opinions of microblogs [29].

Although research into the use of templates as a means of guiding summaries has spanned a variety of domains, research considering hybrid solutions involving templates is underrepresented for tasks including scientific articles.

III. Methods and Implementation

For the creation of the summary, the process was split into multiple stages. This included the creation of a test dataset due to the need for specific metadata and reference summaries of the journal issues as found in the editorials.

The test dataset was made up of 7 issues of “The Journal of Universal Computer Science”, totaling 39 papers. The number of articles per issue varied, as seen in Table I. Each of the papers was pre-processed using GROBID and selected data extracted and saved in JSON format.

TABLE I
Overview of the number of articles contained in the issues, with issues being coded in the form of "volume/number"

Issue	26/07	26/09	26/10	26/11	27/01	28/03	28/10
# articles	4	9	4	8	3	6	5

In the next step, it was necessary to evaluate existing approaches that utilize language models. The summary created in this first stage presents the informational core that is later used to fill the templates. The quality of these subsummaries directly influences the quality of the final issue summary.

For the selection of the models, several conditions were formulated:

- The evaluated models are trained - and later tested - on scientific articles.
- The summary length is one or two sentences, with the result reflecting the overall topic of the article.

- Full sentence summaries are preferred to text fragments only.
- Abstractive single-document summarization

Several models that matched the requirements were evaluated for their performance. As each document was summarized by itself, the focus was on single-document abstractive summarization. The ones considered were LexRank [4], SciTLDR [31], Samsum [32], Pegasus-Pubmed [33], and LongT5 [34]. SciTLDR was used in three different ways: SciTLDR-F used the full text of the article to create the summary, SciTLDR-A used only the abstract, and SciTLDR-AC used the abstract and conclusion.

For an automatic evaluation of readability and complexity, the Python library textstat was used, in particular the Flesch reading ease score [35] and the automated readability index [36]. The Flesch reading ease score typically goes from (below) zero to 100, where lower scores signify higher difficulty, and higher scores easier texts. The automated readability index allocates a grade level to the text, with decimal numbers placing the text in-between two levels. A score of 14 is considered to indicate college-level literature.

Although these scores are typically used to assess the readability of longer literary texts, the choice was made to use them for the selection of the summarization model with the (much shorter) automatically created summaries. As they do not require a reference text to compare against (unlike ROUGE scores), their use was meant to give an indication of text quality and help with the selection of a promising model that returns an easily readable summary. The calculated scores are listed in Table II.

TABLE II
SELECTED LANGUAGE MODELS AND THEIR EVALUATION SCORES IN COMPARISON TO THE REFERENCE SUMMARY

Method	Metric			
	Flesch	Readability	ROUGE-1	ROUGE-L
Reference summary	25.20	20.28	-	-
SciTLDR-F	14.98	25.25	0.6402	0.4707
SciTLDR-A	11.79	26.10	0.6994	0.5713
SciTLDR-AC	6.66	26.91	0.6915	0.5667
LexRank	10.47	29.37	0.5343	0.3756
LongT5	40.89	13.87	0.2337	0.1710
Pegasus-Pubmed	44.33	14.24	0.2393	0.1786
Samsum	28.74	18.35	0.5026	0.4065
T5-one-line	14.78	24.68	0.6794	0.5244

In addition, both ROUGE-1 and ROUGE-L scores were calculated by comparing the automatically created sample summaries to the reference summary, to obtain further information on the performance of the model for this task. Taking into account the different evaluation metrics, the final decision was made according to the ROUGE scores. Flesch and Readability scores are intended to grade readability of text; as shorter sentences are usually considered to have better readability, partial summaries, although not matching the requirements, tended to perform better for these metrics.

The results showed a strong variation, with ROUGE-L scores between 0.1525 and 0.5556. SciTLDR [31], in particular SciTLDR-A, which used only the abstract as input, was found to work best for the intended purpose. Due to the intended short length of the summary, even creating one-sentence summaries leads to generally well-formed, informative results.

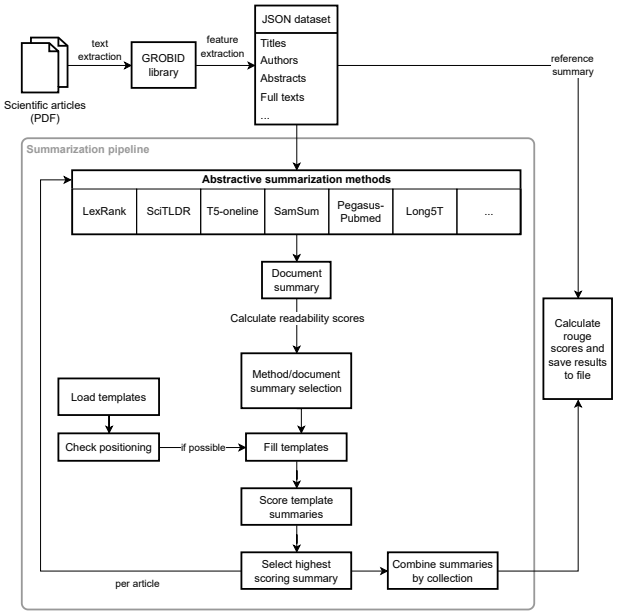

Figure 1. System architecture from PDFs to the final file containing all issue summaries.

The editorials are created using three general steps, with the process visualized in Figure 1: short document summary creation through an abstractive method, the use of templates to complete the summaries according to the pattern present in existing editorials, and post-processing through the use of the natural language toolkit.

The templates are selected according to specified criteria. Each template contains a placement array of the form [x, y, z], where each letter can take the value of 0 or 1, respectively. As an example, the template file of the sentence “Finally, in "[TITLE]", [AUTHORS] [SUMMARY].” contains the placement array [0,0,1], which means that it cannot be placed at the beginning or in the middle of a text. The only valid placement is at the end. This process is included as “Check positioning” in Figure 1. After non-applicable templates are discarded, candidate summaries are created using all remaining templates by combining them with the previously created short summary. These candidate template summaries are scored using the python package *textstat*’s method *text_standard* to evaluate readability. The candidate summary with the highest readability is then selected and combined with all other paper summaries from a specific journal issue.

The resulting summaries are evaluated both by automated means and manually by experts. For automatic evaluation, both ROUGE-1 and ROUGE-L are used.

For manual evaluation, two user studies were conducted for manual evaluation by experts. Each of them placed focus on a different aspect. The first compared the created summaries of

articles with their equivalents from the reference summary and asked the evaluating person to choose the one they preferred overall.

For the second survey, participants received the abstract of an article and its respective automatically created summary. A five-point Likert scale was created to investigate how metrics such as fluency, informativeness, coverage, and accuracy were rated when put in context of the article's abstract.

The combination of both evaluations allowed insight into user preferences and aspects of particular focus.

IV. Results

When scoring the issue summaries using ROUGE-1 and ROUGE-L, the results showed an occasional strong variance, as visible in Table III. One particularly high score was an outlier, while there is no exceptionally low score. Overall, the results are promising but do suggest that it is necessary to pay particular attention to ensuring higher consistency in results to avoid outliers in any direction - though particularly lower scores.

TABLE III
ROUGE scores for each issue, comparing automatic issue summary to manually created reference summary.

Issue	ROUGE-1	ROUGE-L
26/07	0.91	0.73
26/09	0.68	0.53
26/10	0.60	0.53
26/11	0.77	0.64
27/01	0.64	0.49
28/03	0.63	0.42
28/10	0.70	0.56

The manual evaluation took place using two user surveys. Although only completed by a small number of participants, the results are important for future research directions and evaluations in which particular strengths and weaknesses of this approach can be found.

The first survey was started by 11 participants, with 8 of them completing it. The second survey was started by 14 participants and completed by 8, as well. In both cases, incomplete survey results were removed from the evaluation.

For the first survey, in which they noted their preference for either the automatically created summary or the manually created one, in 11 cases, the automatically created summary was preferred. In three cases, the votes were split equally between the two choices. In one notable case, all participants agreed and preferred the manual summary. Upon closer inspection, the automatically created summary was not grammatically correct.

In the second survey, 10 questions asked participants to rate each of the automatically created summaries on four metrics. The overall results were promising, though with a high standard deviation for coverage, fluency, and informativeness, as can be seen in Table IV. Optional free-text answers were given in a minority of cases, but allowed insight into the differing opinions of users that influenced the ratings positively as well as negatively.

TABLE IV
Performance evaluation for each of the given metrics, as well as average score and standard deviation (where "Very Poor" is 1 and "Excellent" is 5

Performance	Accuracy	Coverage	Fluency	Informativeness
Excellent	18	14	23	17
Good	48	35	28	32
Fair	11	13	18	19
Poor	2	16	10	10
Very Poor	1	2	1	2
Average	4	3.54	3.78	3.65
Std. Dev.	0.76	1.07	1.04	1.03

V. Discussion

The results did not suggest a relationship between the number of articles/subsummaries and the ROUGE score calculated for the overall issue summary. For example, both the highest and lowest ROUGE-1 scores were reached by issues that contained 4 articles (26/07 and 26/10). The highest ROUGE-L score was also scored by 26/07, with the lowest being 26/03, made up of 6 articles. Both 26/09 and 26/10 scored close ROUGE-L scores, with the first containing 9 articles and the second containing 4. Therefore, it does not appear that there is significant correlation to be found.

The survey invitations were sent to people in the academic field at a variety of levels of education, from bachelor's students to professors. The answers given - in particular the free-text answers in survey 2 - mirror the different levels of expectations the participants have for scientific summaries. While some participants paid particular attention to how fluently readable a summary was ("The repetition of full names is entirely irrelevant. It makes the sentences VERY hard to read[...]", others paid detailed attention to the wording and commonly used phrases ("Nice! Though it is a run-on sentence. May need a period there to separate it [...]"). Depending on the summary, participants either preferred summaries that were less detailed and more readable ("#2 gives more information but without any context it's hard to understand, #1 is more general" or preferred more detail ("The second summary is more detailed and fits better to the abstract", "Both summaries are of high quality, but #1 just seems to offer a more rounded and comprehensive snapshot of the abstract [...]").

Overall, the average performance of the summary was rated between "Fair" to "Good"; however, it becomes clear that the process is not reliable enough with respect to its output. Although most results are acceptable, there are instances where the summarization process fails to produce a grammatically correct sentence. In the test data, this was the case with one subsummary. In direct comparison to the manually created reference summary, all survey participants considered the automatically created summary sentence inadequate and preferred the reference summary.

The following summaries of a paper included in the test data illustrates this issue [37]:

"Damjan Fujs, Simon Vrhovec and Damjan Vavpotič present

"Bibliometric Mapping of Research on User Training for Secure Use of Information Systems", which conducted bibliometric mapping of research on user training for secure use of information systems."

This summary of the paper was automatically created. It is apparent that the first half of the sentence does not fit well with the second, as it appears that the article itself conducted the mapping instead of the authors. In comparison to the following summary, which was written by the issue's editors in the editorial, the automatically created summary clearly fails to measure up.

"In their paper "Bibliometric Mapping of Research on User Training for Secure Use of Information Systems, Damjan Fujs, Simon Vrhovec and Damjan Vavpotič conduct a bibliometric mapping of research on user training for secure use of information systems [38]."

For use in science, it is thus necessary to further extend or modify the approach explained in this paper to ensure correct grammar of summaries and consistent text quality, as anything less is likely to leave behind unsatisfied users.

VI. Conclusion and Future Work

This paper described an approach for the automatic summarization of scientific articles to create topic overviews. The combination of templates and language models led to results that were overall promising. Two user studies allowed showed where participants found strengths and weaknesses in the automatically created summaries, both compared to a manually created alternative, and when evaluated for specific metrics. The significant standard deviation in score indicates that the target audience should be strongly considered when creating a system such as this. Furthermore, the use of templates is problematic in combination with full sentences that do not necessarily follow a specific grammatical structure. As visible in one result, if a summary sentence is created that was not considered during the creation of the templates, it may lead to grammatically incorrect results that negatively impact the user experience.

Future work may consider the dynamic creation of templates, such as the use of a language model to create a larger variety than is feasible by hand. This would also solve the issue of repetitive sentence structures.

Furthermore, more language models should be considered for use in the future. Due to the constant development in the field, new models constantly appear. It may also be of interest to fine-tune an own model for either the creation of single-sentence summaries or templates.

Finally, it may be useful to increase each sub-summary length according to user preference. A user study may be useful to find the preferred summary length for specific use-cases, in which case a system that allows dynamic selection of sub-summary lengths might be a promising approach.

References

[1] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable AI methods - a brief overview", in *xxAI - Beyond Explainable AI*, Springer International Publishing, 2022, pp. 13–38. DOI: 10.1007/978-3-031-04083-2_2.

[2] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. L. Wang, "MSˆ2: A dataset for multi-document summarization of medical studies", *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7494–7513, 2021. DOI: 10.18653/v1/2021.emnlp-main.594.

[3] H. P. Luhn, "The automatic creation of literature abstracts", *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958. DOI: 10.1147/rd.22.0159.

[4] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization", *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, 2004. DOI: 10.1613/jair.1523.

[5] S. Beliga, A. Meštrović, and S. Martincic-Ipsic, "An overview of graph-based keyword extraction methods and approaches", *Journal of Information and Organizational Sciences*, vol. 39, pp. 1–20, 2015.

[6] S. M. Harabagiu and F. Lacatusu, "Generating single and multi-document summaries with gistexter", in *Document Understanding Conferences*, 2002, pp. 11–12.

[7] Y. Han, F. Li, K. Liu, and L. Liu, "Template based chinese news event summarization", in *Proceedings of the Second International Conference on Semantics, Knowledge, and Grid (SKG'06)*, IEEE, 2006. DOI: 10.1109/skg.2006.102.

[8] T. Oya, Y. Mehdad, G. Carenini, and R. Ng, "A template-based abstractive meeting summarization: Leveraging summary and source text relationships", in *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, Association for Computational Linguistics, 2014, pp. 45–53. DOI: 10.3115/v1/w14-4407.

[9] P. G. Desai, H. Sarojadevi, and N. N. Chiplunkar, "A Template Based Algorithm for Automatic Summarization and Dialogue Management for Text Documents", *International Journal of Research in Engineering and Technology*, vol. 04, no. 11, pp. 334–340, 2015. DOI: 10.15623/ijret.2015.0411059.

[10] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang, "Template-based named entity recognition using BART", in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, 2021, pp. 1835–1845. DOI: 10.18653/v1/2021.findings-acl.161.

[11] Y.-H. Chen, P.-Y. Chen, H.-H. Shuai, and W.-C. Peng, "TemPEST: Soft template-based personalized EDM subject generation through collaborative summarization", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 7538–7545, 2020. DOI: 10.1609/aaai.v34i05.6252.

[12] K. Wang, X. Quan, and R. Wang, "BiSET: Bi-directional selective encoding with template for abstractive summarization", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2019, pp. 2153–2162. DOI: 10.18653/v1/P19-1207.

[13] Z. Cao, W. Li, S. Li, and F. Wei, "Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2018, pp. 152–161. DOI: 10.18653/v1/p18-1015.

[14] A. Vaswani *et al.*, "Attention is all you need", in *Advances in Neural Information Processing Systems*, vol. 30, Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2017, pp. 5999–6009.

[15] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training",

2018, [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (visited on 08/29/2024).

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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*, vol. 1, Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.

[17] Z. Ji *et al.*, "Survey of hallucination in natural language generation", *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023. DOI: 10.1145/3571730.

[18] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, and I. Gurevych, "Ranking generated summaries by correctness: An interesting but challenging application for natural language inference", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2019, pp. 2214–2220. DOI: 10.18653/v1/p19-1213.

[19] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173.

[20] H. Wang, Y. Gao, Y. Bai, M. Lapata, and H. Huang, "Exploring explainable selection to control abstractive summarization", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, pp. 13 933–13 941, 2021. DOI: 10.1609/aaai.v35i15.17641.

[21] S. Jusoh, "A study on nlp applications and ambiguity problems", *Journal of Theoretical & Applied Information Technology*, vol. 96, no. 6, pp. 1486–1499, 2018.

[22] B. P. Yap, A. Koh, and E. S. Chng, "Adapting BERT for word sense disambiguation with gloss selection objective and example sentences", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, 2020, pp. 41–46. DOI: 10.18653/v1/2020.findings-emnlp.4.

[23] D. Patel, S. Shah, and H. Chhinkaniwala, "Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique", *Expert Systems with Applications*, vol. 134, pp. 167–177, 2019. DOI: 10.1016/j.eswa.2019.05.045.

[24] P. Verma and H. Om, "MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization", *Expert Systems with Applications*, vol. 120, pp. 43–56, 2019. DOI: 10.1016/j.eswa.2018.11.022.

[25] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, "Gender bias in neural natural language processing", in *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*. Springer International Publishing, 2020, pp. 189–202. DOI: 10.1007/978-3-030-62077-6_14.

[26] T. Spinde *et al.*, "Automated identification of bias inducing words in news articles using linguistic and context-oriented features", *Information Processing & Management*, vol. 58, no. 3, 2021. DOI: 10.1016/j.ipm.2021.102505.

[27] C. van der Lee, E. Krahmer, and S. Wubben, "Automated learning of templates for data-to-text generation: Comparing rule-based, statistical and neural methods", in *Proceedings of the 11th International Conference on Natural Language Generation*, Tilburg University, The Netherlands: Association for Computational Linguistics, 2018, pp. 35–45. DOI: 10.18653/v1/W18-6504.

[28] J. Sun, Y. Wang, and Z. Li, "An improved template representation-based transformer for abstractive text summarization", in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020. DOI: 10.1109/ijcnn48605.2020.9207609.

[29] I. M. Bilal *et al.*, "Template-based abstractive microblog opinion summarization", *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1229–1248, 2022. DOI: 10.1162/tacl_a_00516.

[30] X. Liu, H. Huang, G. Shi, and B. Wang, "Dynamic prefix-tuning for generative template-based event extraction", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2022, pp. 5216–5228. DOI: 10.18653/v1/2022.acl-long.358.

[31] I. Cachola, K. Lo, A. Cohan, and D. Weld, "TLDR: Extreme summarization of scientific documents", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, 2020, pp. 4766–4777. DOI: 10.18653/v1/2020.findings-emnlp.428.

[32] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization", in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, Association for Computational Linguistics, 2019, pp. 70–79. DOI: 10.18653/v1/d19-5409.

[33] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization", in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 11 328–11 339.

[34] M. Guo *et al.*, "LongT5: Efficient text-to-text transformer for long sequences", in *Findings of the Association for Computational Linguistics: NAACL 2022*, Association for Computational Linguistics, 2022, pp. 724–736. DOI: 10.18653/v1/2022.findings-naacl.55.

[35] R. Flesch, "A new readability yardstick.", *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[36] G. Thomas, R. D. Hartley, and J. P. Kincaid, "Test-retest and inter-analyst reliability of the automated readability index, flesch reading ease score, and the fog count", *Journal of Reading Behavior*, vol. 7, no. 2, pp. 149–154, 1975.

[37] D. Fujs, S. Vrhovec, and D. Vavpotič, "Bibliometric mapping of research on user training for secure use of information systems", *JUCS - Journal of Universal Computer Science*, vol. 26, no. 7, pp. 764–782, 2020, ISSN: 0948-695X. DOI: 10.3897/jucs.2020.042. eprint: https://doi.org/10.3897/jucs.2020.042.

[38] S. Wendzel *et al.*, "Information security methodology, replication studies and information security education", *JUCS - Journal of Universal Computer Science*, vol. 26, no. 7, pp. 762–763, 2020, ISSN: 0948-695X. DOI: 10.3897/jucs.2020.041. eprint: https://doi.org/10.3897/jucs.2020.041.