Objective Argument Summarization in Search

Timon Ziegenbein[1,†] Shahbaz Syed[2,†] Martin Potthast[3] Henning Wachsmuth[1]

[1] Leibniz University Hannover
{t.ziegenbein, h.wachsmuth}@ai.uni-hannover.de
[2] Leipzig University
shahbaz.syed@uni-leipzig.de
[3] University of Kassel, hessian.AI, and ScaDS.AI
martin.potthast@uni-kassel.de

Abstract. Decision-making and opinion formation are influenced by arguments from various online sources, including social media, web publishers, and, not least, the search engines used to retrieve them. However, many, if not most, arguments on the web are informal, especially in online discussions or on personal pages. They can be long and unstructured, subjective and emotional, and contain inappropriate language. This makes it difficult to find relevant arguments efficiently. We hypothesize that, on search engine results pages, "objective snippets" of arguments are better suited than the commonly used extractive snippets and develop corresponding methods for two important tasks: *snippet generation* and *neutralization*. For each of these tasks, we investigate two approaches based on (1) prompt engineering for large language models (LLMs), and (2) supervised models trained on existing datasets. We find that a supervised summarization model outperforms zero-shot summarization with LLMs for snippet generation. For neutralization, using reinforcement learning to align an LLM with human preferences for suitable arguments leads to the best results. Both tasks are complementary, and their combination leads to the best snippets of arguments according to automatic and human evaluation.

Keywords: Computational Argumentation · Information Retrieval · Large Language Models · Text Summarization · Text Neutralization

1 Introduction

Deliberative processes are a key element of well-informed decision-making and opinion formation. Their goal is to explore and evaluate the space of arguments that are relevant for deciding on the best course of action in a given situation [34]. Vast amounts of arguments on virtually all topics of interest can be found on the web and are retrievable using generic or specialized search engines. However, the argument snippets returned by argument search engines are often insufficient to help users find relevant arguments—for two main reasons. First, the standard methods for generating snippets often fail to capture the essence of an argument [2]

[†] Equal contribution.

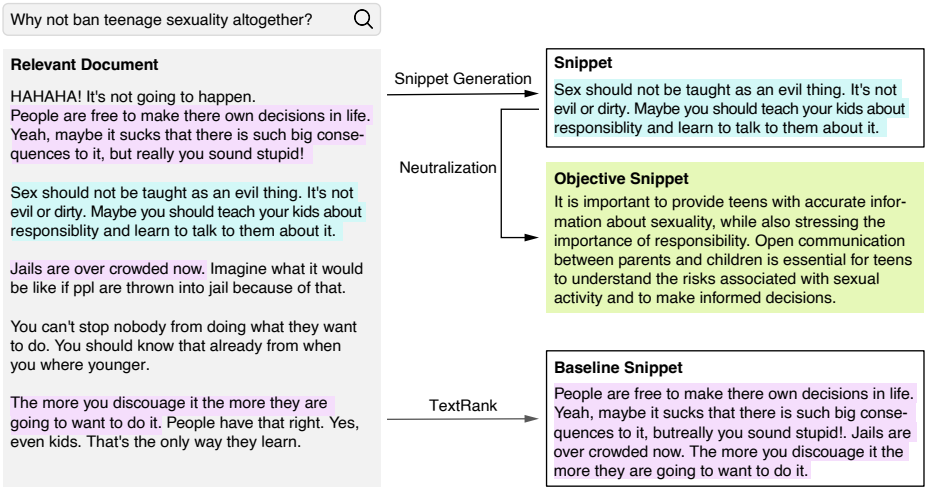

Fig. 1. Illustration of our two-step approach encompassing snippet generation and neutralization to create an objective snippet of a relevant document (argumentative text) for a user query (controversial issue). The document contains information that is relevant to the query, although written inappropriately. Our objective snippet mitigates this while retaining the relevant content. For comparison, an extractive TextRank baseline reflects this inappropriateness, resulting in a potentially ineffective snippet.

(henceforth referred to as the argument's "gist"). Second, the snippets often contain subjective, informal, emotional, or inappropriate language that distracts from the gist [38]. Though the original arguments may still contain information that is highly relevant to a topic, snippets that reflect inappropriate presentations may prevent users from recognizing them as relevant.

In this paper, we investigate whether "objective snippets" are better suited for argument search engines. We define such a snippet to combine the main claim of an argument and the evidence supporting it (basically, the gist), while avoiding overly subjective and informal language. We propose a two-step approach to create objective snippets of arguments. The first step, snippet generation, aims to extract the main message and supporting evidence of an argument. We assume that a short summary of an argument (i.e. two sentences) can represent this gist. The second step, neutralization, aims to neutralize the language of the extracted core statement to make it more objective. We also investigate the necessity of neutralization as a separate task, since abstractive summaries in particular can potentially neutralize the language of the source text already during generation. Figure 1 exemplifies snippets from existing snippet generation models as well as from our approach. It demonstrates that existing approaches produce snippets that retain inappropriate language, which undermines the effectiveness of the main argument. In contrast, our approach combines snippet generation and

neutralization to produce objective, well-written snippets while preserving the semantics of the source argument. Our contributions are as follows:[4]

- A two-step approach that tackles the tasks of snippet generation and neutralization to create objective argument snippets for argument search (Section 3).
- Three manual evaluation studies on snippet generation and neutralization, individually and in combination, using (1) the args.me corpus [1] and (2) the appropriateness corpus [38] as ground truth (Sections 4 and 5).

We show that abstractive snippets are better suited to present arguments as search results than extractive snippets. In particular, argument neutralization leads to an expected increase in the likelihood of a productive discussion on the topic. Moreover, combining abstractive summarization with neutralization creates a more objective snippet that further improves the already-preferred abstractive snippets in terms of the likelihood that users are willing to read the full argument presented by the snippet.

2 Related Work

In this section, we describe relevant previous work on the tasks of snippet generation and neutralization. Since snippet generation is very similar to summarization, we describe relevant work from both areas.

2.1 Snippet Generation

Snippets in search engines are primarily extractive in nature. Snippet generators extract the most relevant parts of the source text, especially those containing the terms of the query [3, 14, 32, 35]. The aim of a snippet is to help users quickly identify documents likely to satisfy their information need [9]. First, argument search engines such as args.me [33] or ArgumenText [28] used the first sentences of retrieved arguments as snippets. Later, extractive snippets of the arguments as proposed by Alshomary et al. [2] replaced them, enriching TextRank with argumentative information to extract the main claim and supporting premise as an argument snippet, which forms a baseline in our evaluation. The arguments were also summarized in individual sentences [28], key points [4], and conclusions [30].

Our motivation is to introduce objective snippets of arguments in a search engine. While minimizing the reuse of text in the snippets (from the source) is beneficial [7], traditionally, extractive summaries are preferred over abstractive summaries to avoid incorrect rephrasing of facts from the source text. This is because abstractive summaries of standard sequence-to-sequence models suffer from hallucinations [24] and incorrectly merge different parts of the source, leading to incorrect facts [5]. However, recent advances in abstractive summarization using pre-trained language models have been shown to generate more fluid and coherent summaries than purely extractive approaches, which improves their

[4]The experiment code is available at https://github.com/webis-de/RATIO-24

overall readability and preference by humans [13]. Therefore, we opt for abstractive snippets in this work. Moreover, we investigate the zero-shot effectiveness of the instruction-driven Alpaca [31] model using prompting.

2.2 Neutralization

Neutralization can be seen as a style transfer task. Style transfer in the context of natural language generation aims to control attributes in the generated text, such as politeness, emotion, or humor among many others [17]. Text style transfer has been applied to authorial features and literary genres [12]. Most studies deal with broad notions of style, including the formality and subjectivity of a text [18]. There are also approaches to changing sentiment polarity (of reviews) [16], political bias (of news headlines) [6], and framing (of news articles) [8].

Many approaches learn a sequence-to-sequence model on parallel source–target text pairs. Modifying the style often works reliably, but preserving the content seems to be a challenge [6]. On the other hand, style and content are difficult to separate in text (i.e., words can reflect both simultaneously). To mitigate this, some works avoid disentangling latent representations of style and content [10], but this cannot guarantee that certain information is preserved. Others restrict transfer to low-level linguistic decisions [12, 27].

Our aim is to improve the appropriateness of arguments to ensure that they are suitable for a wide audience. However, unlike traditional style transfer, the role of semantic preservation here is rather superficial, as some parts of our texts that are responsible for inappropriateness may be inappropriate due to their content rather than their style, such as ad hominem attacks. Therefore, we generally prefer appropriateness over semantic similarity in this paper.[5] Since no parallel data is available for the argument neutralization task, we rely on an instruction-based zero-shot approach with Alpaca [31]. For further refinement, we use the appropriateness classifier from Ziegenbein et al. [38] and an adapted version of the RLHF (Reinforcement Learning using Human Feedback) method from Stiennon et al. [29]. The authors of Madanagopal and Caverlee [23] use a reinforcement learning-based approach to correct subjective language in Wikipedia articles, which comes closest to our approach. However, their approach is based on parralel data, which is not available for the task of neutralization. As far as we know, there is no style transfer approach for argument neutralization to date, and none of the related reinforcement learning approaches for style transfer use prompting as the initial model (i.e., for the policy).

3 Approach

This section describes the approaches we evaluated for generating argument snippets and their neutralization.

[5]The role of semantic similarity is being investigated in another paper under review.

3.1 Snippet Generation

We investigated three snippet generation approaches: (1) an unsupervised extractive argument summarization model, (2) a supervised abstractive news summarization model, and (3) an instruction-tuned zero-shot summarization model.

Extractive-Summarizer. With TextRank, Alshomary et al. [2] proposed an unsupervised extractive argument snippet generation approach that extracts the main claim and premise of an argument as its snippet. To identify the corresponding argument sentences, a variant of PageRank [26] is used to rank them based on their contextual importance and argumentativeness. Starting from equal scores for all sentences, the model iteratively updates these scores until convergence is achieved. The two highest-scoring sentences are then extracted in their original order to maintain coherence. TextRank serves as the standard model for generating snippets for the args.me search engine and as our baseline.

Abstractive-Summarizer. For supervised snippet generation, we use a BART model [21], finetuned to the task of abstractive news summarization on the CNN/DailyMail dataset [25].[6] To tailor its summaries to the task argument snippet generation, we shorten the input to 102 tokens and limit the minimum and maximum summary length to 25% and 35% of the argument length respectively.

Instruction-Summarizer. To instruct Alpaca to generate a snippet, we use the prompt `### Instruction: The following is an argument on the topic "<topic>". Extract a coherent gist from it that is exactly two sentences long. ### Input: <argument> ### Response:` and insert an argument and its topic. Generation is done at a temperature of 1 and sampling with a p-value of 0.95. The number of generated sentences is limited to two in order to ensure snippets of a similar length compared to the other approaches.

3.2 Neutralization

For neutralization, we compare (1) an instruction-tuned zero-shot neutralization model, and (2) a reinforcement learning-aligned neutralization model.

Instruction-Neutralizer. To instruct Alpaca to neutralize a text, we use the prompt `### Instruction: Rewrite the following argument on the topic of "<topic>" to be more appropriate and make only minimal changes to the original argument. ### Input: <argument> ### Response:` and provide it with the argument and its topic. We use a temperature of 1 and sample with a p-value of 0.95 during generation. The number of generated tokens is limited to 50% to 150% of the original argument to ensure that the model does not delete or add too much content when rewriting the arguments or snippets.

Aligned-Neutralizer. To align Alpaca with human-defined appropriateness criteria, we finetune it using reinforcement learning from human feedback [29, 39]. During the training process, we use the same prompt settings and hyperparameters as before, but adjust the output of the model to generate texts that are categorized as appropriate by the appropriateness classifier of Ziegenbein et al. [38]. Thus,

[6] https://huggingface.co/facebook/bart-large-cnn

texts generated by Alpaca serve as input to the classifier and the returned probability value for the appropriateness class as a reward to update Alpaca. For efficiency, we do not update Alpaca's original weights but use adapter-based low-rank adaptation (LoRA) [15]. A full description of the approach and the training process is part of a paper soon to be published [37].[7]

4 Data

For evaluation, we use two datasets sampled from (1) the args.me corpus [1] and (2) the appropriateness corpus [38]. The former is used to evaluate the snippet generation approaches and combining snippet generation and neutralization, while the latter is used to evaluate the argument neutralization approaches.

4.1 The args.me Corpus

To obtain the dataset for our snippet generation experiments, we sample arguments from the args.me corpus [1]. The args.me corpus contains 387,606 arguments from four debate portals, each annotated with a stance (pro or con) and a topic (e.g., "abortion" or "gay marriage"). Based on the ten most frequently submitted queries to the args.me API [33], we created an initial dataset. To ensure adequate summarization potential for snippet generation and to account for possible input length limitations of the models used in our experiments, we filter the dataset to contain only arguments between 100 and 500 words in length. Furthermore, we use an ensemble classifier based on the five folds of the appropriateness corpus to retain only inappropriate arguments. Finally, we extract the top five pro and top five con arguments for each query based on the args.me ranking obtained from its API. This gives us a final dataset of 99 arguments.[8]

4.2 The Appropriateness Corpus

To obtain the dataset for our neutralization experiments, we sample arguments from the appropriateness corpus [38]. The corpus contains 2,191 arguments labeled with the corresponding discussion titles from three genres (reviews, discussion forums, and Q&A forums). Each argument is annotated by three annotators according to a 14-dimensional taxonomy of inappropriateness errors. We filter the corpus to include only arguments that were classified as inappropriate by all three annotators in the original study to ensure that there is a clear need for neutralization. As before, we only retain arguments between 100 and 500 words in length. Finally, we draw a random sample of 100 arguments from the corpus to obtain our final dataset.

[7] The code and data used to train the models can be found here: https://github.com/timonziegenbein/appropriateness-style-transfer

[8] As one of the queries did not contain enough arguments to meet the inappropriateness criteria, one query contains only nine arguments instead of ten.

Table 1. Evaluation of the snippet generation approaches without neutralization: (a) ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), and BERTScore (Sim.), computed between the source argument and the generated snippet, perplexity (PPL) of the generated snippet and percentage of appropriate generated snippets (App.). (b) Absolute counts of ranks assigned by human evaluators to the three approaches and their average.

Approach	**(a) Automatic**						**(b) Manual**			
	R1	R2	RL	Sim.	PPL↓	App.↑	#1	#2	#3	Avg.↓
Extractive-Sum.	0.29	0.28	0.29	0.25	67.7	0.21	42	126	**327**	2.58
Abstractive-Sum.	**0.40**	**0.38**	**0.38**	**0.35**	50.9	0.31	**274**	149	72	**1.59**
Instruction-Sum.	0.24	0.11	0.16	0.13	**26.5**	**0.58**	179	**220**	96	1.83

5 Evaluation

We evaluate our approaches in a series of experiments, both automatically and manually. For automatic evaluation, we quantify the content preservation of all approaches with ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) [22] for lexical similarity, and with BERTScore (Sim.) [36] for semantic similarity. Furthermore, we measure the fluency of the generated texts with Perplexity (PPL) and compute the percentage of instances for which an approach was able to change the label from inappropriate to appropriate (based on the ensemble classifier of Ziegenbein et al. [38], see Section 3). The manual evaluation is detailed in the corresponding subsections, as the user studies differ for each of the tasks.

5.1 Snippet Generation

Automatic Evaluation. Table 1a shows that, when automatically determining the best summarization model for snippet generation, the Abstractive-Summarizer scores best in terms of content preservation (highest R1, R2, RL, Sim.). Instruction-Summarizer is strongest in fluency (PPL 26.5) and creates appropriate snippets for 58% of inappropriate arguments. The extractive baseline Extractive-Summarizer does not win in any of the automatic measurements used.

Manual Evaluation. We hired five evaluators on upwork.com who are native English speakers and tasked them to evaluate snippets of 99 arguments from our three models: Instruction-Summarizer, Abstractive-Summarizer, and Extractive-Summarizer. Given a topic, a source argument (pro/con) and three snippets, the evaluators rated the suitability of a snippet to be displayed on a search engine results page for the argument by ranking them from "best" to "worst." A detailed annotation guide describing the characteristics of a good snippet, such as high coverage of key information from the original argument and its ability to help users easily identify relevant arguments from a ranking of results.

As shown in Table 1b, Abstractive-Summarizer proved to be the best model for generating snippets according to the evaluators, ranking first in about 56% of the examples (274 out of 495). The agreement between annotators was 0.22,

Table 2. Evaluation of the neutralization approaches: (a) ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BERTScore (Sim.), perplexity (PPL) of the neutralized argument, and percentage of successfully neutralized arguments (App.). (b) Absolute counts of ranks assigned by the human evaluators to the three approaches and their average.

Approach	**(a) Automatic**						**(b) Manual**			
	R1	R2	RL	Sim.	PPL↓	App.↑	#1	#2	#3	Avg.↓
Exact-Copy	**1.00**	**1.00**	**1.00**	**1.00**	66.1	0.00	10	91	**399**	2.78
Instruction-Neut.	0.79	0.66	0.73	0.67	29.5	0.40	67	**345**	88	2.04
Aligned-Neut.	0.41	0.16	0.27	0.18	**18.4**	**0.97**	**423**	64	13	**1.18**

as measured by Kendall's τ rank correlation coefficient [19]. This indicates a positive rank correlation while underlining the subjectivity of the quality ratings.

5.2 Neutralization

Automatic Evaluation. Comparing the Instruction-Neutralizer with the Aligned-Neutralizer, Table 2a shows that there are differences in content preservation and transfer of appropriateness. That is, the Instruction-Neutralizer performs better on R1 (0.79), R2 (0.66), RL (0.73), and Sim. (0.67), whereas the Aligned-Neutralizer performs better on fluency (PPL 18.4) and transfer (App. 0.97), making almost all arguments appropriate (97%). This suggests that there is a trade-off between retaining the content of the argument and improving appropriateness. As mentioned above, we are investigating this effect in another paper that is not yet published at the time of writing. However, a manual inspection of the neutralized arguments and our annotators' comments shows that, despite the rather low content preservation (0.18 for BERTScore), the main meaning of the argument and its reasoning are mostly preserved, but the arguments do not show any lexical similarity to the original argument.

Manual Evaluation. If people prefer neutralized arguments over the baseline arguments that contain inappropriate content, this is evidence that neutralization is useful for the ultimate goal of creating "objective snippets." Accordingly, we evaluated the neutralized arguments of Instruction-Neutralizer and Aligned-Neutralizer together with the baseline argument. Like above, five human evaluators ranked the three argument variants from "best" to "worst" according to their appropriateness to be presented in a civil debate on a given topic. We used 100 (manually labeled) inappropriate arguments from the appropriateness corpus. The evaluators were provided with a comprehensive guide describing the characteristics of inappropriate arguments and how to identify them [38].

Table 2b shows the results. Neutralized arguments from Aligned-Neutralizer are preferred over others in 84.6% of cases (423 out of 500). This underlines the effectiveness of neutralization and its implicit goal of making arguments more appropriate in public debates. Kendall's τ for this evaluation was 0.48, indicating a positive correlation between the rankings. Compared to the snippet generation

Table 3. Evaluation of the combined approach (snippet generation + neutralization): (a) ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BERTScore (Sim.), perplexity (PPL) of the generated snippet, and percentage of appropriate snippets generated (App.). (b) Absolute and relative count of snippets of one approach being preferred over the other.

Approach	**(a) Automatic**						**(b) Manual**	
	R1	R2	RL	Sim.	PPL↓	App.↑	Pref.↑	%↑
Abstractive-Sum.	**0.40**	**0.38**	**0.38**	**0.35**	50.9	0.31	57	0.11
+ Aligned-Neut.	0.25	0.10	0.17	0.11	**20.0**	**0.87**	**438**	**0.89**

task, the evaluators were able to distinguish more reliably between the quality of inappropriate and appropriate variants of an argument.

5.3 Objective Snippets

Automatic Evaluation. Comparing the two approaches using our automatic measures, Table 3a shows that combining Abstractive-Summarizer with Aligned-Neutralizer further decreases the similarity of the snippet to the original argument (0.35 vs. 0.11), but increases the number of appropriate snippets (0.87 vs. 0.31).

Manual Evaluation. In addition to evaluating the individual subtasks, we also evaluated the holistic approach by assessing the usefulness of the objective snippets. Specifically, we performed a pairwise comparison between the objective snippets and the non-neutralized snippets. In contrast to evaluating the generation of the snippets, where the original argument was also provided, we only provided the topic to the five human evaluators. Given a self-contained query, they were asked to select the excerpt they were most likely to click on to read the full argument. For this evaluation, we used 100 arguments for 10 topics from the args.me corpus and selected an equal number of pro and con arguments.

Table 3b shows the results. Objective snippets were preferred over non-neutralized snippets in 89% of the cases (438 out of 495). This indicates that neutralization has a positive effect on the likelihood that search engine users will follow the link to read the full argument from which the snippet was extracted. Krippendorff's α [20] was 0.29, indicating moderate agreement between annotators. Further examples of snippets generated by our best approach (Abstractive-Sum. + Aligned-Neut.) are shown in Table 5 in the Appendix.

Qualitative Analysis. We conducted a manual evaluation of each task, which included the generation and neutralization of snippets as well as the resulting objective snippets generated with our approach. For all tasks, we recruited annotators who are native English speakers, aiming for a balanced representation of male and female annotators. Annotators had the opportunity to provide comments and could also contact us directly if they needed help. No additional questions were asked throughout the annotation tasks, with the exception of a brief review of a small subset of completed annotations to confirm understanding of the task.

Table 4. Quality dimensions for each tasks (snippet generation, neutralization, objective snippets), derived from the comments of annotators in our manual evaluation studies.

Task	Quality Dimensions (Preferred by Annotators)
Snippet Generation	specificity, clarity, positive/inoffensive language, conciseness, self-containment, informativeness, focus on the issue, avoiding personal attacks, structure and coherence, accuracy/correctness
Neutralization	openness, simple language, absence of profanity, facilitating critical evaluation, seriousness, absence of grammatical/orthographic errors, balanced emotions, well-reasoned, structure and coherence, formal language, non-speculative
Objective Snippets	conciseness, simple language, fluency, balanced emotions, includes quotes/evidence/statistics, specificity, coherence

For each example within our three studies, annotators were asked to provide optional feedback in natural language on their ratings and preferences for the results of each study. We manually analyzed nearly 500 comments to identify important quality dimensions for achieving the goal of creating objective snippets. In particular, we derived quality dimensions that have been studied in related areas such as summarization, text generation, and sentiment analysis. Table 4 provides an overview of these dimensions for each task. Examples of comments for the tasks of snippet generation, neutralization, and objective snippets are shown in Tables 6 and 7 in the Appendix, respectively.

Overall, we found that grammaticality and positive language strongly influenced the credibility and acceptability of the argument snippets. Annotators consistently preferred arguments that were free of spelling errors, had correct punctuation, and were well-structured, regardless of their content. Therefore, ensuring grammatical correctness and a well-structured output is crucial. Furthermore, the use of positive language is preferred over negative language, with annotators emphasizing that a positive tone signals critical thinking and openness to other opinions. Consequently, neutralization plays a key role in ensuring that the snippets are suitable for a wide audience. In line with the quality dimensions of summaries [11], high-quality annotators preferred snippets that were informative, concise and coherent.

Limitations and Ethical Concerns

This paper aims to provide evidence that objective argument snippets significantly improve the overall user experience when searching for arguments. While our human annotators strongly advocate neutralizing arguments and their snippets, we currently lack evidence that directly correlates (to a large extent) with satisfying users' information needs. Another unexplored aspect is to investigate whether the generation of snippets, especially through prompting, implicitly

incorporates neutralization to some extent. These questions are subject to future research in the given context.

It is crucial to note that the success of generating and neutralizing snippets is closely linked to the quality of the original arguments. In cases where the original arguments are poorly constructed or unclear, the resulting objective snippets may not effectively represent their gist. We also recognize that neutralization is not appropriate in certain contexts where preserving the original language of the source text is critical (e.g., student essays, legal documents, or medical fields). In such cases, the application of neutralization requires the user's consent to ensure transparency and accountability. Practical implementations of our approach could include user options that allow individuals to choose between the original and neutralized versions of a snippet or an argument. We further acknowledge that our assumption that the generated arguments are gists of the original arguments may not always hold true. In some cases, the generated arguments may not capture the essence of the original arguments, leading to a loss of information.

We would like to acknowledge that the task of creating and neutralizing snippets is to a certain extent subjective. The choice of the best snippet may vary depending on the annotator's background, experience, and personal preferences. For this reason, we believe that further research is needed to explore the influence of these factors on the quality of the generated snippets and, in particular, to involve the authors of the original arguments in the process of snippet generation.

In summary, our empirical research highlights the potential benefits of mitigating subjective bias, particularly in the broader context of engaging with the opinions and arguments of others. This does not only facilitate informed decision making, but it can also be valuable for educational purposes.

6 Conclusion

In this paper, we have investigated the hypothesis that "objective snippets" of arguments are better for argument search engine results than state-of-the-art extractive snippets, using methods that combine snippet generation and neutralization. Our study has conveyed that a BART-based supervised summarization model outperforms a zero-shot Alpaca model to snippet generation. For neutralization, we have found that using reinforcement learning to align a large language model with human preferences for suitable arguments works best. We have also observed that both tasks complement each other and that their combination leads to the most effective snippets, as shown by human evaluation. Our results provide important insights and innovative methods that can be used to improve search engines in order to produce more efficient search results for users.

Acknowledgment

This project has been partially funded by the German Research Foundation (DFG) within the project OASiS, project number 455913891, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999), and by the EU, as part of the project OpenWebSearch.eu, grant agreement number 101070014, as part of the Horizon Europe research and innovation program.

Bibliography

[1] Ajjour, Y., Wachsmuth, H., Kiesel, J., Potthast, M., Hagen, M., Stein, B.: Data acquisition for argument search: The args.me corpus. In: KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings, pp. 48–59 (2019), https://doi.org/10.1007/978-3-030-30179-8_4, URL https://doi.org/10.1007/978-3-030-30179-8_4

[2] Alshomary, M., Düsterhus, N., Wachsmuth, H.: Extractive snippet generation for arguments. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pp. 1969–1972, ACM (2020), https://doi.org/10.1145/3397271.3401186, URL https://doi.org/10.1145/3397271.3401186

[3] Bando, L.L., Scholer, F., Turpin, A.: Constructing query-biased summaries: a comparison of human and system generated snippets. In: Proceedings of the third symposium on Information interaction in context, pp. 195–204 (2010)

[4] Bar-Haim, R., Eden, L., Friedman, R., Kantor, Y., Lahav, D., Slonim, N.: From arguments to key points: Towards automatic argument summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 4029–4039, Association for Computational Linguistics (2020), URL https://www.aclweb.org/anthology/2020.acl-main.371/

[5] Cao, Z., Wei, F., Li, W., Li, S.: Faithful to the original: Fact aware neural abstractive summarization. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pp. 4784–4791, AAAI Press (2018), URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16121

[6] Chen, P., Wu, F., Wang, T., Ding, W.: A semantic qa-based approach for text summarization evaluation. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pp. 4800–4807, AAAI Press (2018), URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16115

[7] Chen, W., Syed, S., Stein, B., Hagen, M., Potthast, M.: Abstractive snippet generation. In: Huang, Y., King, I., Liu, T., van Steen, M. (eds.) WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, pp. 1309–1319, ACM / IW3C2 (2020), https://doi.org/10.1145/3366423.3380206, URL https://doi.org/10.1145/3366423.3380206

[8] Chen, W.F., Al Khatib, K., Stein, B., Wachsmuth, H.: Controlled neural sentence-level reframing of news articles. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2683–2693, Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021), https://doi.org/10.18653/v1/2021.findings-emnlp.228, URL https://aclanthology.org/2021.findings-emnlp.228
[9] Croft, W.B., Metzler, D., Strohman, T.: Search engines: Information retrieval in practice, vol. 520. Addison-Wesley Reading (2010)
[10] Dai, N., Liang, J., Qiu, X., Huang, X.: Style Transformer: Unpaired text style transfer without disentangled latent representation. arXiv:1905.05621 [cs] (2019)
[11] Dang, H.T.: Overview of duc 2005. In: Proceedings of the document understanding conference, vol. 2005, pp. 1–12 (2005)
[12] Gero, K., Kedzie, C., Reeve, J., Chilton, L.: Low level linguistic controls for style transfer and content preservation. In: Proc. of the 12th Int. Conference on Natural Language Generation, pp. 208–218 (2019)
[13] Goyal, T., Li, J.J., Durrett, G.: News summarization and evaluation in the era of GPT-3. CoRR **abs/2209.12356** (2022), https://doi.org/10.48550/arXiv.2209.12356, URL https://doi.org/10.48550/arXiv.2209.12356
[14] Groeneveld, D., Meyerzon, D., Mowatt, D.: Generating search result summaries (2010), uS Patent 7,853,587
[15] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net (2022), URL https://openreview.net/forum?id=nZeVKeeFYf9
[16] Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., Xing, E.P.: Toward controlled generation of text. In: Proc. of the 34th Int. Conference on Machine Learning, vol. 70, pp. 1587–1596 (2017)
[17] Jin, D., Jin, Z., Hu, Z., Vechtomova, O., Mihalcea, R.: Deep learning for text style transfer: A survey. Comput. Linguistics (1), 155–205 (2022), https://doi.org/10.1162/COLI_A_00426, URL https://doi.org/10.1162/coli_a_00426
[18] Kabbara, J., Cheung, J.C.K.: Stylistic transfer in natural language generation systems using recurrent neural networks. In: Proc. of the Workshop on Uphill Battles in Language Processing, pp. 43–47 (2016)
[19] Kendall, M.G.: Rank correlation methods. (1948)
[20] Krippendorff, K.: Estimating the reliability, systematic error and random error of interval data. Educational and Psychological Measurement (1), 61–70 (1970)
[21] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: 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, ACL 2020, Online, July 5-10, 2020,

pp. 7871–7880, Association for Computational Linguistics (2020), URL https://www.aclweb.org/anthology/2020.acl-main.703/

[22] Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81, Association for Computational Linguistics, Barcelona, Spain (Jul 2004), URL https://www.aclweb.org/anthology/W04-1013

[23] Madanagopal, K., Caverlee, J.: Reinforced sequence training based subjective bias correction. In: Vlachos, A., Augenstein, I. (eds.) Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2585–2598, Association for Computational Linguistics, Dubrovnik, Croatia (May 2023), https://doi.org/10.18653/v1/2023.eacl-main.189, URL https://aclanthology.org/2023.eacl-main.189

[24] Maynez, J., Narayan, S., Bohnet, B., McDonald, R.T.: On faithfulness and factuality in abstractive summarization. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 1906–1919, Association for Computational Linguistics (2020), URL https://doi.org/10.18653/v1/2020.acl-main.173

[25] Nallapati, R., Zhou, B., dos Santos, C.N., Gülçehre, Ç., Xiang, B.: Abstractive text summarization using sequence-to-sequence rnns and beyond. In: Goldberg, Y., Riezler, S. (eds.) Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016, pp. 280–290, ACL (2016), https://doi.org/10.18653/v1/k16-1028, URL https://doi.org/10.18653/v1/k16-1028

[26] Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)

[27] Pryzant, R., Martinez, R.D., Dass, N., Kurohashi, S., Jurafsky, D., Yang, D.: Automatically neutralizing subjective bias in text. In: Proc. of the Thirty-Fourth AAAI Conference on Artificial Intelligence, pp. 480–489 (2020)

[28] Stab, C., Daxenberger, J., Stahlhut, C., Miller, T., Schiller, B., Tauchmann, C., Eger, S., Gurevych, I.: ArgumenText: Searching for arguments in heterogeneous sources. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 21–25, Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://doi.org/10.18653/v1/N18-5005, URL https://www.aclweb.org/anthology/N18-5005

[29] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.: Learning to summarize from human feedback. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Curran Associates Inc., Red Hook, NY, USA (2020), ISBN 9781713829546

[30] Syed, S., Khatib, K.A., Alshomary, M., Wachsmuth, H., Potthast, M.: Generating informative conclusions for argumentative texts. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, Findings of

ACL, vol. ACL/IJCNLP 2021, pp. 3482–3493, Association for Computational Linguistics (2021), https://doi.org/10.18653/v1/2021.findings-acl.306, URL https://doi.org/10.18653/v1/2021.findings-acl.306

[31] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model (2023), URL https://github.com/tatsu-lab/stanford_alpaca

[32] Tombros, A., Sanderson, M.: Advantages of Query Biased Summaries in Information Retrieval. In: Proceedings of SIGIR 1998, pp. 2–10 (1998)

[33] Wachsmuth, H., Potthast, M., Al-Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., Stein, B.: Building an argument search engine for the web. In: Proceedings of the 4th Workshop on Argument Mining, pp. 49–59, Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017), https://doi.org/10.18653/v1/W17-5106, URL https://www.aclweb.org/anthology/W17-5106

[34] Walker, M.: Information and deliberation in discourse. In: Intentionality and Structure in Discourse Relations (1993), URL https://aclanthology.org/W93-0238

[35] White, R., Ruthven, I., Jose, J.M.: Web document summarisation: a task-oriented evaluation. In: 12th International Workshop on Database and Expert Systems Applications, pp. 951–955, IEEE (2001)

[36] Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, Proceedings of Machine Learning Research, vol. 119, pp. 11328–11339, PMLR (2020), URL http://proceedings.mlr.press/v119/zhang20ae.html

[37] Ziegenbein, T., Skitalinskaya, G., Makou, A.B., Wachsmuth, H.: LLM-based Rewriting of Inappropriate Argumentation using Reinforcement Learning (2024)

[38] Ziegenbein, T., Syed, S., Lange, F., Potthast, M., Wachsmuth, H.: Modeling Appropriate Language in Argumentation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4344–4363, Association for Computational Linguistics (Jul 2023), https://doi.org/10.18653/v1/2023.acl-long.238, URL https://aclanthology.org/2023.acl-long.238.pdf

[39] Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P.F., Irving, G.: Fine-tuning language models from human preferences. CoRR (2019), URL http://arxiv.org/abs/1909.08593

A Appendix

Table 5. Examples of arguments from the appropriateness corpus and their neutralized snippets generated by our approach.

Type	**Example**
Argument	I always hear people talking about how the death penalty needs to be enforced more often. I, personally, am undecided on my opinion on the death penalty, but in my mind, the death penalty is used primarily as a societal form of retaliation (much the way that putting someone's head on a pike was used in the past). [...] At the very least, it will disseminate information to public about what execution in the United States is actually like (this may, in turn, change the beliefs that someone holds on the death penalty).
Neutralized Snippet	It is important for the public to be aware of the death penalty in the U.S. in order to make informed decisions about the practice. Televising executions could help to bridge the gap in understanding between the public and the criminal justice system.
Argument	At Poland Spring, we know that freshness and consistency go hand in hand. Our rigorous quality control standards help us make sure that you will enjoy the same crisp, delicious water every time you take a sip. To maintain the purity of our water, we employ a comprehensive, multi-barrier filtration system that involves carefully controlled and continuously monitored disinfection processes in hygienically designed lines. [...] Every year, we screen for over 200 possible contaminants-far more than state and federal regulations require.
Neutralized Snippet	Poland Spring utilizes a rigorous filtration system, which is tested regularly throughout the production process to ensure the safety of the water. This process ensures the water quality remains consistent and meets strict quality standards.
Argument	CarCraze — you are wrong, sorry to say What you have there, my dear, is a bona fide Court Order for you to appear in court to answer to a Motion to Collect a Debt. [...] If you "just let it ride", as you say, don't be surprised when you open up your paycheck and you find a good chunk of it GONE ! What you should do: Seek legal counsel (get a lawyer) and come up with a manageable payment plan that is both satisfactory to your creditor, and workable for you. DON'T HESITATE - you will regret it.
Neutralized Snippet	If you choose to ignore the notice of motion from the collection firm, it is possible that you may incur additional fees and charges. It is best to seek legal counsel from a professional in order to create a manageable payment plan that suits your financial situation.
Argument	In my opinion, all it serves to do is remind men to fulfil the traditional gender roles. To "be a man" you have to provide for your family, be responsible and protect your family. It's the male equivalent of "be more ladylike". All it accomplishes is: * Make non-traditionally masculine men feel like shit * [...] And god forbid that men have some values that women are 'supposed' to have. Why not have a gender neutral phrase? "Be a better person"?
Neutralized Snippet	This phrase can be harmful as it can lead to feelings of exclusion for men who don't conform to traditional gender roles, as these expectations can be seen as exclusive to men. This can be damaging as it may make those men feel like they do not belong.

Table 6. A sample of comments provided by annotators organized by quality dimensions that influence snippet generation. Comments are edited for presentation purposes. Also, the anonymized snippets referred to in the comments (e.g., A, B, C) do not always correspond to a specific model being evaluated as the order of the snippets was randomized.

Dimension	**Annotators' Comments**
Conciseness	"This snippet provides a concise and clear definition of feminism, emphasizing equality and respect for both men and women."
Focus	"Snippet A is ranked highest for its clear emphasis on the necessity of abortion... "
Offensive Language	"While this snippet discusses activism against woman abuse and negative elements related to women, it introduces terms that may be considered offensive (e.g., "SLuts")."; "The use of language like "I am going to have to ask you go to timeout because that idea is downright childish" might be perceived as confrontational."
Informativeness	The first snippet condenses the argument very succinctly and covers most of the major points in the arguments above.
Structure & Coherence	"Snippet B is the worst summary for the argument presented above because there is not direct link between the statement and the conclusion of the snippet - so it completely misses the point."; "Snippet C is by far the best snippet in this sequence. It has a clear structure and it delivers the message of the paragraph."
Grammaticality	"Snippet A and B both have grammatical errors (need/needed), which would discredit the link/argument/page from the get go."
Self-contained	"Snippet A is ranked 3rd because there is no logical link between the first part of the snippet and the second part of the snippet. No reader could understand what the argument is about from that summary alone."
Accuracy	"Snippet C comes in last because it is completely inaccurate, given that it claims these points of view are Trump's points of view. In fact, they are the views of the narrator/author."
Argument-friendly Vocabulary	"Argument A has slightly more argument-friendly vocabulary (e.g ' juxtaposition' used in contrast to 'antithesis')."; "The only main difference between argument A and C is the choice of vocabulary to describe the couples that the writer is associated with being either committed or monogamous. I think that the use of the word 'committed' to describe the couple in argument A makes the example used more relevant to the argument at hand."
Seriousness	"Although argument A and C are similar, argument A has a more sincere tone and slightly more proper grammar: e.g Latinos not "Latinos" / Latin America not Latinamerica."
Profanity & Speculation	"B is more to the point, doesn't speculate on strategies and has no profanity like C/B (Shit/assholes)."
Clarity	"This argument is clear in its meaning, provides a concise comparison between the two cases, and avoids inappropriate language or tone." , "This argument presents the issue clearly, maintains a proportional and balanced perspective by addressing both sides..."

Table 7. A sample of comments provided by annotators organized by quality dimensions that influence preference of snippets. Comments are edited for presentation purposes. Also, the anonymized arguments referred to in the comments (e.g., A, B, C) do not always correspond to a specific model being evaluated as the order of the arguments was randomized.

Dimension	**Annotators' Comments**
Respectful	"It begins with a dismissive tone ("is totally crap")...", "This argument is the best as it presents its points in a clear and respectful manner.", "This argument uses sarcasm ("(pause here for deeply bitter laugh)") and refers to a political figure in a dismissive manner ("The Shrub").", "This argument is the most appropriate as it maintains a professional tone, focuses on the key issues, and promotes a respectful and balanced discussion of the pro-choice vs. pro-life debate."
Critical Evaluation	"The mention of "corrupt the minds of my children" is emotionally charged, which may not provide room for critical evaluation.", "the ending part "expecting male users to do the looking for both themselves and the women" may come off as slightly dismissive, which makes it less open to others' arguments."
Formal Language	"This argument and Argument A are quite similar, but Argument C uses slightly more refined and formal language, making it more appropriate for a professional discussion. For example, it uses "Fourth Amendment" instead of "4th" and "naive" instead of "living under a rock"."
Grammaticality	"It has orthographic errors (e.g., missing spaces and inconsistent punctuation), making it harder to follow. Some of the phrasing is repetitive, and its presentation can hinder a clear understanding of its main points.", "...contains orthographic errors ("ur", "shld", "dugs"), and uses casual and unclear language. This decreases its credibility and appropriateness for a professional debate."
Conciseness	"I chose snippet A because it uses short sentences instead of one long one, and because it uses numbers which is more concrete than just saying "a high degree."", "Both very similar but snippet B is more concise..."
Evidence	"Snippet B is very subjective and doesn't present any evidence for the argument.", "...uses more numbers which encourages me to read.", "Snippet B is less pushy and provides more examples to back up its argument."
Grammaticality	"Snippet B has grammar and spelling errors which discourages me from wanting to read more."
Critical Evaluation	"A places the onus of thought on the reader, allowing them the space to form their own opinions. B is instructional, seemingly saying everything that is needed for a reader to make their mind up without their own research."
Structure & Coherence	"Snippet A clearly outlines their argument, while snippet B hops back and forth from one point to another without a linear thought process."