Claim Optimization in Computational Argumentation

Gabriella Skitalinskaya[1,2], Maximilian Spliethöver[1], and Henning Wachsmuth[1]
[1]Leibniz University Hannover, Institute of Artificial Intelligence
[2]University of Bremen, Department of Computer Science
{g.skitalinska,m.spliethoever,h.wachsmuth}@ai.uni-hannover.de

Abstract

An optimal delivery of arguments is key to persuasion in any debate, both for humans and for AI systems. This requires the use of clear and fluent claims relevant to the given debate. Prior work has studied the automatic assessment of argument quality extensively. Yet, no approach actually improves the quality so far. To fill this gap, this paper proposes the task of *claim optimization*: to rewrite argumentative claims in order to optimize their delivery. As multiple types of optimization are possible, we approach this task by first generating a diverse set of candidate claims using a large language model, such as BART, taking into account contextual information. Then, the best candidate is selected using various quality metrics. In automatic and human evaluation on an English-language corpus, our quality-based candidate selection outperforms several baselines, improving 60% of all claims (worsening 16% only). Follow-up analyses reveal that, beyond copy editing, our approach often specifies claims with details, whereas it adds less evidence than humans do. Moreover, its capabilities generalize well to other domains, such as instructional texts.

1 Introduction

The delivery of arguments in clear and appropriate language is a decisive factor in achieving persuasion in any debating situation, known as *elocutio* in Aristotle's rhetoric (El Baff et al., 2019). Accordingly, the claims composed in an argument should not only be grammatically fluent and relevant to the given debate topic, but also unambiguous, self-contained, and more. Written arguments therefore often undergo multiple revisions in which various aspects are optimized (Zhang and Litman, 2015).

Extensive research has been done on the automatic assessment of argument quality and the use of large language models on various text editing tasks. Yet, no work so far has studied how to actually improve argumentative texts. However, developing respective approaches is a critical step towards building effective writing assistants, which could help learners write better argumentative texts (Wambsganss et al., 2021) or rephrase arguments made by an AI debater (Slonim et al., 2021). In this work, we close the outlined gap by studying how to employ language models for rewriting argumentative text to optimize its delivery.

Debate topic	Should humans be allowed to explore DIY gene editing?
Previous claim	Humans should be allowed to explore DIY gene editing.
Original claim	This technology could be weaponized.
Optimized claim 1	This technology could be weaponized and harmful to human beings.
Optimized claim 2	This technology could be used by criminals to create and weaponize bio-mechanisms.
Optimized claim 3	This technology could be weaponized, so it is important to safeguard it from being weaponized.

Figure 1: Examples of different optimized versions of an *original claim* found on the debate platform Kialo. All optimizations were generated by the approach proposed in this paper, using the *debate topic* as context.

We start by defining the task of *claim optimization* in Section 3, and adjust the English-language claim revision dataset of Skitalinskaya et al. (2021) for evaluation. The new task requires complementary abilities: On the one hand, different types of quality issues inside a claim must be detected, from grammatical errors to missing details. If not all quality aspects can be improved simultaneously, specific ones must be targeted. On the other hand, improved claim parts need to be integrated with the context of the surrounding discussion, while preserving the original meaning as far as possible. Figure 1 shows three exemplary optimizations of a claim from the debate platform *Kialo*. The first elaborates what the consequence of weaponization

is, whereas the second rephrases the claim to clarify what weaponizing means, employing knowledge about the debate topic. The third renders the stance of the claim explicit. We observe that different ways to optimize a claim exist, yet the level of improvement differs as well.

To account for the multiplicity of claim optimization, we propose a controlled generation approach that combines the capabilities of large language models with quality assessment (Section 4). First, a fine-tuned generation model produces several candidate optimizations of a given claim. To optimize claims, we condition the model on discourse context, namely the debate topic and the previous claim in the debate. The key to selecting the best optimization is to then score candidates using three quality metrics: *grammatical fluency*, *meaning preservation*, and *argument quality*. Such candidate selection remains understudied in many generative tasks, particularly within computational argumentation.

In automatic and manual evaluation (Section 5), we demonstrate the effectiveness of our approach, employing fine-tuned BART (Lewis et al., 2020) for candidate generation. Our results stress the benefits of quality assessment (Section 6). Incorporating context turns out especially helpful for making shorter claims—where the topic of the debate is difficult to infer—more self-contained. According to human annotators, our approach improves 60% of all claims and harms only 16%, clearly outperforming standard fine-tuned generation.

To gain further insights, we carry out a manual annotation of 600 claim optimizations and identify eight types typically found in online debate communities, such as *elaboration* and *disambiguation* (Section 7). Intriguingly, our approach covers similar optimization types as in human revisions, but we also observe limitations (Section 7). To explore to what extent it generalizes to other revision domains, we also carry out experiments on instructional texts (Anthonio and Roth, 2020) and formal texts (Du et al., 2022), finding that it outperforms strong baselines and state-of-the-art approaches.

In summary, the contributions of this paper are:

1. *a new task*, claim optimization, along with a manual analysis of typical optimization types;
2. *a computational approach* that selects the best generated candidate claim in terms of quality;
3. *empirical insights* into the impact and challenges of optimizing claims computationally.[1]

2 Related Work

Quality assessment has become a key topic in computational argumentation research (Lapesa et al., 2023). Various quality dimensions exist in argumentation theory, as surveyed by Wachsmuth et al. (2017) and assessed computationally in various works (Lauscher et al., 2020; Marro et al., 2022). Many of them relate to quality aspects we consider in this work, from clarity and organization (Wachsmuth et al., 2016) to the general evaluability of arguments (Park and Cardie, 2018), potential fallacies in their reasoning (Goffredo et al., 2022), and the appropriateness of the language used (Ziegenbein et al., 2023). Recently, (Skitalinskaya and Wachsmuth, 2023) tackled the question whether an argumentative claim is in need of revision, whereas Jundi et al. (2023) investigated where to best elaborate a discussion. While Gurcke et al. (2021) leverage claim generation for a refined assessment of argument quality, we are not aware of any prior work that actually optimizes arguments or their components in order to improve quality.

As shown in Figure 1, there can be several ways to optimize a given text. Our key idea is to select the best optimization among diverse candidates generated by a language model. Prior generation work on candidate selection hints at the potential benefits of such setup, albeit in other tasks and domains. In early work on rule-based conversational systems, Walker et al. (2001) introduced dialogue quality metrics to optimize template-based systems towards user satisfaction. Kondadadi et al. (2013) and Cao et al. (2018) chose the best templates for generation, and Mizumoto and Matsumoto (2016) used syntactic features to rank candidates in grammar correction. Recently, Yoshimura et al. (2020) proposed a reference-less metric trained on manual evaluations of grammar correction system outputs to assess generated candidates, while Suzgun et al. (2022) utilize pre-trained language models to select the best candidate in textual style transfer tasks.

In generation research on computational argumentation, candidate selection remains largely understudied. Most relevant in this regard is the approach of Chakrabarty et al. (2021) which reframes arguments to be more trustworthy (e.g., less partisan). It generates multiple candidates and selects

[1] Data, code, and models from our experiments are found at https://github.com/GabriellaSky/claim_optimization

one based on the entailment relation scores to the input. Extending this idea, we select candidates based on various properties, including argument quality.

Understanding the editing process of arguments is crucial, as it reveals what quality dimensions are considered important. For Wikipedia, Daxenberger and Gurevych (2013) proposed a fine-grained taxonomy as a result of their multi-label edit categorization of revisions (Daxenberger and Gurevych, 2012). The taxonomy focuses solely on the editing actions performed, such as inserting, deleting, and paraphrasing. In contrast, Yang et al. (2017) identified various semantic intentions behind Wikipedia revisions, from *copy editing* to *content clarifications* and *fact updates*. Their taxonomy defines a starting point for our research. Not all covered intentions generalize beyond Wiki scenarios, though.

Wikipedia-based corpora have often been used in the study of editing and rewriting, including paraphrasing (Max and Wisniewski, 2010), grammar correction (Lichtarge et al., 2019), bias neutralization (Pryzant et al., 2020), and controllable text editing (Faltings et al., 2021; Du et al., 2022). Similarly, WikiHow enabled summarization (Koupaee and Wang, 2018) and knowledge acquisition (Zhou et al., 2019). However, neither of these includes *argumentative* texts. Instead, we thus rely on the corpus of Skitalinskaya et al. (2021), which consists of revision histories of argumentative claims from online debates. Whereas the authors *compare* claims in terms of quality, we propose and study the new task of automatically *optimizing* claim quality. Moreoever, we see the revision types they distinguish (clarification, grammar correction, linking to external sources) as too coarse-grained to represent the diversity of claim optimizations. We refine them manually into eight optimization types, allowing for a more systematic analysis. Skitalinskaya et al. (2021) also found low correlations between the revision types and 15 common argument quality dimensions (Wachsmuth et al., 2017), suggesting that they are rather complementary. Primarily, they target the general form a well-phrased claim should have and its relevance to the debate.

For the analysis of argumentative text rewriting, Zhang and Litman (2015) incorporated both argumentative writing features and surface changes. To explore the classification of essay revisions, they defined a two-dimensional schema, combining the revision operation (e.g., modify, add, or delete) with the component being revised (e.g., reasoning or evidence). Moreover, Afrin and Litman (2018) created a small corpus of between-draft revisions of 60 student essays to study whether revision improves quality. However, these works do not uncover the reasoning behind a revision operation and are more geared towards analysis at the essay level.

3 Task and Data

This section introduces the proposed task and presents the data used for development and evaluation.

3.1 Claim Optimization

We define the claim optimization task as follows:

Task Given as input an argumentative claim c, potentially along with context information on the debate, rewrite c into an output claim $\tilde{c}$ such that

(a) $\tilde{c}$ improves upon c in terms of text quality and/or argument quality, and

(b) $\tilde{c}$ preserves the meaning of c as far as possible.

While we conceptually assume that c consists of one or more sentences and has at least one quality flaw, our approaches do not model this explicitly. Moreover, note that c might have multiple flaws, resulting in $n \geq 2$ candidate optimizations $\tilde{C} = \{\tilde{c}_1, \ldots, \tilde{c}_n\}$. In this case, the goal is to identify the candidate $c^* \in \tilde{C}$ that maximizes overall quality.

3.2 Data for Development and Evaluation

We start from the ClaimRev dataset (Skitalinskaya et al., 2021), consisting of 124,312 claim revision histories from the debate platform *Kialo*. Each history defines a chain $(c_1, ..., c_m)$, in which claim c_i is a revised version of the previous claim, c_{i-1} with $1 < i \leq m$, improving upon its quality. According to the authors, this holds in 93% of all cases.

From each revision chain, we derived all possible optimization pairs $(c, \tilde{c}) := (c_{i-1}, c_i)$, in total 210,222. Most revisions are labeled with their intention by the users who performed them, rendering them suitable for learning to optimize claims automatically.[2] Overall, 95% of all pairs refer to three intention labels: *clarification*, *typo/grammar correction*, and *corrected/added links*. To avoid noise from the few remaining labels, we condensed the data to 198,089 instances of the three main labels.[3]

[2] As 26% of all pairs were unlabeled, we trained a BERT model to assign such pairs one of the 6 most prominent labels.

[3] The labels of the removed instances denote changes to the meaning of c and statements from which no action or intention can be derived (e.g., "see comments", "moved as pro").

For the final task dataset, we associated each remaining pair $(c, \tilde{c})$ to its context: the *debate topic* τ (i.e., the thesis on Kialo) as well as the *previous claim* $\hat{c}$ (the parent on Kialo), which is supported or opposed by c (see Figure 1). We sampled 600 revision pairs pseudo-randomly as a test set (200 per intention label), and split remaining pairs into training (90%) and validation set (10%). As the given labels are rather coarse-grained, we look into the optimizations in more detail in Section 7.

4 Approach

We now present the first approach to automatic claim optimization. To account for the variety of possible optimizations, multiple candidate claims are generated that are pertinent to the context given and preserve the claim's meaning. Then, the best candidate is selected based on quality metrics. Both steps are detailed below and illustrated in Figure 2.

4.1 Seq2Seq-based Candidate Generation

To generate candidates, we fine-tune a Seq2Seq model on pairs $(c, \tilde{c})$, by treating the original claim c as encoder source and revised claim $\tilde{c}$ as the decoder target. In a separate experiment, we condition the model on context information, the debate topic τ and the previous claim $\hat{c}$, during fine-tuning to further optimize the relevance of generated candidates. The context is separated from c by delimiter tokens (Keskar et al., 2019; Schiller et al., 2021).

Multiple ways to improve c exist, especially if it suffers from multiple flaws, since not all flaws may be fixed in a single revision. Therefore, we first generate n suitable candidates, $\tilde{c}_1, \ldots, \tilde{c}_n$, among which the best one is to be found later (n is set to 10 in Section 5). However, the top candidates created by language models often tend to be very similar. To increase the diversity of candidates, we perform top-k sampling (Fan et al., 2018), where we first generate the most probable claim (top-1) and then vary k with in steps of 5 (e.g. top-5, top-10, etc).

4.2 Quality-based Candidate Selection

Among the n candidates, we aim to find the optimal claim, c^*, that most improves the delivery of c in terms of text and argument quality. Similar to Yoshimura et al. (2020), we tackle this task as a candidate selection problem. In our proposed strategy, *AutoScore*, we integrate three metrics: (1) grammatical fluency, (2) meaning preservation, and (3) argument quality. This way, we can *explicitly* favor

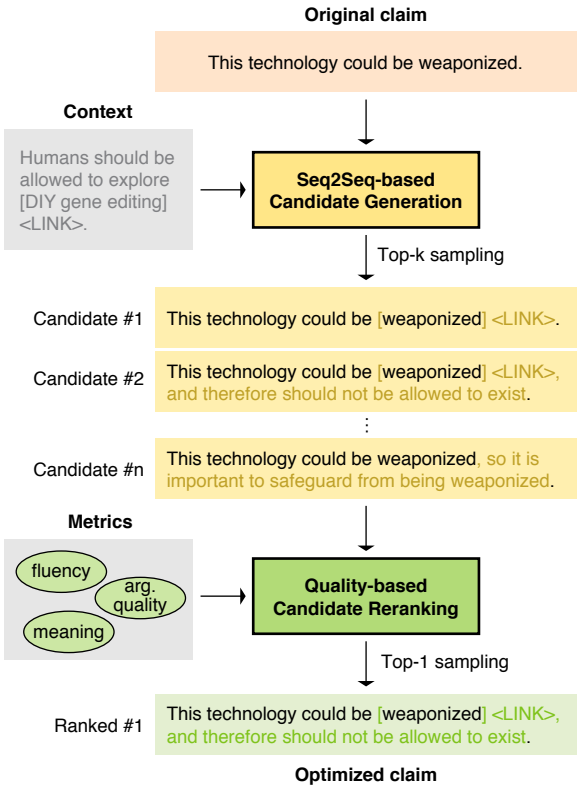

Figure 2: Proposed claim optimization approach: First, we generate n candidates from the *original claim*, possibly conditioned on context information. Then, the *optimized claim* is selected using three quality metrics.

specific quality dimensions via respective models:

Grammatical Fluency We learn to assess fluency on the MSR corpus (Toutanova et al., 2016) where the grammaticality of abstractive compressions is scored by 3–5 annotators from 1 (disfluent) to 3 (fluent). We chose this corpus, since multiple compressions per input make a trained model sensitive to the differences in variants of a text. For training, we average all annotator scores and make the task binary, namely, a text is seen as disfluent unless all annotators gave score 3. Then, we train BERT on the data to obtain fluency probabilities (details found in Appendix A). The accuracy of our model on the suggested data split is 77.4.

Meaning Preservation To quantify to what extent a generated candidate maintains the meaning of the original claim, we compute their semantic similarity as the cosine similarity of the SBERT sentence embeddings (Reimers and Gurevych, 2019).

Argument Quality Finally, to examine whether the generated candidates are better than the original claim from an argumentation perspective, we fine-tune a BERT model on the task of pairwise

argument classification using the ClaimRev dataset. Since this corpus is also used to fine-tune the Seq2Seq model, we apply the same training and validation split as described in Section 3.2 to avoid data leakage, and obtain 75.5 accuracy. We then use its probability scores to determine relative quality improvement (for more details see Appendix A).

Given the three quality metrics, we calculate the final evaluation score, $AutoScore$, as the weighted linear sum of all three individual scores as

$$\alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument,$$

where $fluency$, $meaning$, and $argument$ are normalized scores of the three outlined quality metrics. The non-negative weights satisfy $\alpha + \beta + \gamma = 1$.

It should be noted that depending on the domain or writing skills of the users, there may be other more suitable datasets or approaches to capturing the outlined quality aspects, which could potentially lead to further performance improvements. While we do explore how well the suggested approaches transfer to certain other domains of text (see Section 7.3), identifying the optimal model for each quality dimension falls beyond the scope of this paper.

5 Experiments

This section describes our experimental setup to study how well the claims from Section 3 can be improved using our approach from Section 4. We focus on the impact of candidate selection.

5.1 Seq2Seq-based Candidate Generation

For candidate generation, we employ the pre-trained conditional language model BART (Lewis et al., 2020), using the *bart-large* checkpoint. However, other Seq2Seq architectures can also be considered within our approach (see Appendices A, B).

5.2 Quality-based Candidate Selection

We evaluate our candidate selection approach in comparison to three ablations and four baselines:

Approach To utilize AutoScore for choosing candidates, the optimal weighting of its metrics must be determined. We follow Yoshimura et al. (2020), performing a grid search in increments of 0.01 in the range of 0.01 to 0.98 for each weight to maximize the Pearson's correlation coefficient between AutoScore and the original order of the revisions from revision histories in the validation set. Similar has been done for counterargument retrieval by Wachsmuth et al. (2018). The best weights found are $\alpha = 0.43$, $\beta = 0.01$, and $\gamma = 0.56$, suggesting that meaning preservation is of low importance and potentially may be omitted. We suppose this is due to the general similarity of the generated candidates, so a strong meaning deviation is unlikely.

Ablations To assess the impact of each considered quality metric used in AutoScore, we perform an ablation study, where optimal candidates are chosen based on the individual metric scores:

- *Max Fluency.* Highest grammatical fluency
- *Max Argument.* Highest argument quality
- *Max Meaning.* Highest semantic similarity

Baselines We test four selection strategies for 10 candidates generated via top-k sampling:

- *Unedited.* Return the original input as output.
- *Top-1.* Return the most likely candidate (obtained by appending the most probable token generated by the model at each time step).
- *Random.* Return candidate pseudo-randomly.
- *SVMRank.* Rerank candidates with SVMRank (Joachims, 2006). Using sentence embeddings we decide which of the claim versions is better, by fine-tuning SBERT (*bert-base-cased*) on the corpus of Skitalinskaya et al. (2021).

5.3 Evaluation

We explore claim optimization on all 600 test cases, both automatically and manually:

Automatic Evaluation We compare all content selection strategies against the reference revisions using the precision-oriented *BLEU* (Papineni et al., 2002), recall-oriented *Rouge-L* (Lin, 2004), *SARI* (Xu et al., 2016), which computes the average F_1-scores of the added, kept, and deleted n-grams in comparison to the ground truth revision output, and the *exact match accuracy*. We also compute the semantic similarity of the optimized claim and the context information to capture whether conditioning claims on context affects their topic relevance.

Manual Evaluation As we fine-tune existing generation models rather than proposing new ones, we focus on the *candidate selection* in two manual annotation studies. For each instance, we acquired five independent crowdworkers via *MTurk*.

In the first study, the annotators scored all candidates with respect to the three considered quality metrics. We used the following Likert scales:

- *Fluency.* 1 (major errors, disfluent), 2 (minor errors), and 3 (fluent)
- *Meaning Preservation.* 1 (entirely different), 2 (substantial differences), 3 (moderate differences), 4 (minor differences), and 5 (identical)
- *Argument Quality.* 1 (notably worse than original), 2 (slightly worse), 3 (same as original), 4 (slightly improved), and 5 (notably improved)

A challenge of crowdsourcing is to ensure good results (Sabou et al., 2014). To account for this, we obtained the fina fluency, argument quality and meaning preservation scores using MACE (Hovy et al., 2013), a Bayesian model that gives more weight to reliable workers. In the given case, 39% of the 46 annotators had a MACE competence value > 0.3, which can be seen as reasonable in MTurk studies.

In the second study, we asked annotators to rank four candidates, returned by the content selection strategies, by perceived overall quality. If multiple candidates were identical, we showed each only once. While Krippendorff's α agreement was only 0.20 and percent agreement was 0.36% (majority voting), such values are common in subjective tasks (Wachsmuth et al., 2017; Alshomary et al., 2021).

6 Results and Discussion

Apart from evaluating the applicability of large generative language models to the task of argumentative claim optimization in general, our experiments focus on two questions: (1) Does the use of explicit knowledge about text and argument quality lead to the selection of better candidates? (2) Does the use of contextual information make the generated candidates more accurate and relevant to the debate?

6.1 Overall Claim Optimization Performance

Automatic Evaluation Table 1 shows the automatic scores of all considered candidate selection strategies. The high scores of the baseline *Unedited* on metrics such as BLEU and ROUGE-L indicate that many claim revisions change little only. In contrast, *Unedited* is worst on SARI, a measure taking into account words that are added, deleted, and kept in changes, making it more suitable for evaluation. Here, *BART+AutoScore* performs best on SARI (43.7) and exact match accuracy (8.3%).

Approach	BLEU	RouL	SARI	NoEd↓	ExM
Baselines					
Unedited	**69.4**	**0.87**	27.9	1.00	0.0%
BART + Top-1	64.0	0.83	39.7	0.31	7.8%
BART + Random	62.6	0.83	38.7	0.28	6.8%
BART + SVMRank	55.7	0.76	38.8	0.03	4.5%
Approach					
BART + AutoScore	59.4	0.80	**43.7**	**0.02**	**8.3%**
Ablation					
BART + Max Fluency	57.6	0.78	41.5	0.09	5.8%
BART + Max Argument	60.9	0.81	43.6	0.02	8.0%
BART + Max Meaning	69.0	0.87	33.8	0.72	5.2%

Table 1: Automatic evaluation: Performance of each candidate selection strategy on 600 test cases in terms of BLEU, Rouge-L, SARI, ratio of unedited cases, and ratio of exact matches to target reference.

Model	Strategy	Fluency	Argument	Meaning	Rank
BART	Top-1	2.29	3.61	3.65	2.16
	Random	2.26	3.50	3.53	2.06
	SVMRank	**2.33**	**3.69**	**3.66**	1.95
	AutoScore	**2.33**	3.61	3.57	**1.92**

Table 2: Manual evaluation: Scores on the 600 test cases generated by BART using our candidate selection strategy *AutoScore* or the baselines: fluency (1–3), argument quality and meaning (1–5), mean rank (1–4, lower better). AutoScore ranks significantly better than *Top-1* ($p < .005$), *Random* ($p < .05$), and *SVMRank* ($p < .1$).

The *BART+Max Meaning* ablation supports the intuition that the candidates with highest meaning preservation scores are those with minimal changes, if any (72% of the candidates remain identical to the input). Such identical outputs are undesirable, as the claims are not optimized successfully, which is also corroborated by the low weight parameter ($\beta = 0.01$) found for the meaning preservation metric when optimizing AutoScore (see Section 5).

Manual Evaluation Table 2 shows that human annotators prefer optimized candidates selected by *AutoScore*, with an average rank of 1.92. The difference to *Top-1* and *Random* is statistically significant ($p < .05$ in both cases) according to a Wilcoxon signed-rank test, whereas the gain over the second-best algorithm, *SVMRank*, is limited. Also, candidates of AutoScore and SVMRank are deemed more fluent than those of Top-1 and Random (2.33 vs. 2.29 and 2.26). In terms of argument quality, the results deviate from the automatic evaluation (Table 1), showing marginally higher scores for SVMRank and Top-1. Further analysis revealed that AutoScore and SVMRank agreed on the optimal candidate in 35% of the cases, partially

Context	BLEU	Original	Previous	Topic
Claim only	59.4	0.95	0.55	0.55
+ Previous Claim	**60.3**	0.95	**0.57**	**0.57**
+ Debate Topic	60.0	0.95	0.55	0.55
Human-Baseline	100.0	0.94	0.55	0.55

Table 3: BLEU and semantic similarity score with respect to the *original* claim, the debate's *previous* claim, and its *topic* of BART+AutoScore, depending on the context given for the 600 test samples.

explaining their close scores. Although SVMRank achieved high scores across the three quality metrics, we note that the annotators preferred candidates scores generated by AutoScore, highlighting the importance of more diverse revision changes reflected by lower meaning preservation scores.

Overall, our findings suggest that using candidate selection approaches that incorporate quality assessments (i.e., AutoScore and SVMRank) leads to candidates of higher fluency and argument quality while preserving the meaning of the original claim. In addition to Figure 1, examples of automatically-generated optimized claims can be found in the appendix.

6.2 Performance with Context Integration

Table 3 shows the semantic similarity of claims optimized by our approach and context information, depending on the context given. The results reveal slight improvements when conditioning the model on the previous claim (e.g., 60.3 vs. 59.4 BLEU). To check whether this led to improved claims, two authors of the paper compared 600 claims generated with and without the use of the previous claim in terms of (a) which claim seems better overall and (b) which seems more grounded. We found that using the previous claim as context improved quality in 12% of the cases and lowered it in 1% only, while leading to more grounded claims in 36%.

Qualitative Analysis Our manual inspection of a claim sample revealed the following insights:

First, conditioning on context reduces the number of erroneous specifications, particularly for very short claims with up to 10 words. This seems intuitive, as such claims often convey little information about the topic of the debate, making inaccurate changes without additional context likely.

Next, Kialo revisions often adhere to the following form: A claim introduces a statement and/or supporting facts, followed by a conclusion. This pattern was frequently mimicked by our approach. Yet, in some cases, it added a follow-up sentence repeating the original claim in different wording or generated conclusions containing fallacious or unsound phrases contradicting the original claim in others. Modeling context mitigated this issue.

Finally, we found that models conditioned on different contexts sometimes generated candidates optimized in different regards, whereas a truly optimal candidate would be a fusion of both suggestions.

7 Analysis

To explore the nature of claim optimization and the capabilities of our approach, this section reports on (a) what types of optimizations exist, (b) how well our approach can operationalize these, and (c) how well it generalizes to non-argumentative domains.

7.1 Taxonomy of Optimization Types

To understand the relationship between optimizations found in the data and the underlying revision intentions, two authors of this paper inspected 600 claim revisions of the test set. Opposed to actions, intentions describe the goal of an edit (e.g., making a text easier to read) rather than referring to specific changes(e.g., paraphrasing or adding punctuation). We build on ideas of Yang et al. (2017) who provide a taxonomy of revision intentions in Wikipedia texts. Claims usually do not come from encyclopedias, but from debate types or from monological arguments, as in essays (Persing and Ng, 2015). Therefore, we adapt the terminology of Yang et al. (2017) to gear it more towards argumentative texts.

As a result of a joint discussion of various sample pairs, we decided to distinguish eight optimization types, as presented in Table 4. Both authors then annotated all 600 test pairs for these types, which led to only 29 disagreement cases, meaning a high agreement of 0.89 in terms of Cohen's κ. These cases were resolved by both annotators together.[4]

Table 4 also shows cooccurrences of the types and intention labels. *Typo/grammar correction* and *correcting/adding links* align well with *copy editing* and *corroboration* respectively. In contrast, clarification is broken into more fine-grained types, where *specification* seems most common with 58

[4]We acknowledge that there is potential bias inherent in self-annotation. However, we would like to point out that no knowledge about the test set was used to develop the approach presented in Section 4.

#	Optimization	Description of the Type	Clarification	Grammar	Links
1	Specification	Specifying or explaining a given fact or meaning (of the argument) by adding an example or discussion without adding new information.	58	1	–
2	Simplification	Removing information or simplifying the sentence structure, e.g., with the intent to reduce the complexity or breadth of the claim.	43	–	–
3	Reframing	Paraphrasing or rephrasing a claim, e.g., with the intent to specify or generalize the claim, or to add clarity.	29	–	–
4	Elaboration	Extending the claim by more information or adding a fact with the intent to make the claim more self-contained, sound, or stronger.	23	–	–
5	Corroboration	Adding, editing, or removing evidence in the form of links that provide supporting information or external resources to the claim.	8	–	153
6	Neutralization	Rewriting a claim using a more encyclopedic or neutral tone, e.g., with the intent to remove bias or biased language.	7	–	–
7	Disambiguation	Reducing ambiguity, e.g., replacing pronouns by concepts mentioned before in the debate, or replacing acronyms with what they stand for.	7	–	1
8	Copy editing	Improving the grammar, spelling, tone, or punctuation of a claim, without changing the main point or meaning.	41	200	52

Table 4: Descriptions of the eight claim optimization types identified in the 600 test pairs. The right columns show the count of claims per type for each of the three intention labels from Skitalinskaya et al. (2021): *clarification*, typo/*grammar* correction, and correcting/adding *links*. Note, that a revision may be assigned to multiple categories.

Type	Human	Approach	Better	Same	Worse
Specification	59	152	65%	19%	16%
Simplification	43	18	61%	28%	11%
Reframing	29	21	62%	33%	5%
Elaboration	23	55	62%	18%	20%
Corroboration	161	38	53%	23%	24%
Neutralization	7	0	–	–	–
Disambiguation	8	8	63%	25%	12%
Copy editing	293	301	59%	26%	15%
Overall	623	593	60%	24%	16%

Table 5: Manual analysis: Comparison of the human-optimized claims of all 600 test cases (some have multiple) and of the claims optimized by BART+AutoScore (15 claims were unchanged). The three right columns show the ratio of optimized claims judged *better*, *same*, or *worse* than the original in terms of overall quality.

cases, followed by *simplification* and *reframing*. Examples of each type are found in the appendix.

We point out that the eight types are not exhaustive for all possible claim quality optimizations, but rather provide insights into the semantic and discourse-related phenomena observed in the data. We see them as complementary to the argument quality taxonomy of Wachsmuth et al. (2017) as ways to improve the delivery-related quality dimensions: *clarity*, *appropriateness*, and *arrangement*.

7.2 Performance across Optimization Types

To enable comparison between the human optimizations and automatically generated outputs, two authors of the paper labeled 600 optimized claims with the types defined in Table 4. Due to resource constrains only the best performing approach, BART+AutoScore, was considered. Overall, our approach generates better claims in 60% of the cases, while 84% remain at least of similar quality.

Most noteworthily, we observe that our approach performs optimizations of the type *specification* 2.5 times as often as humans, and more than double as many *elaboration* revisions (55 vs. 23). In contrast, it adds, edits, or removes evidence in the form of links (*corroboration*) four times less often than humans. The model also made fewer *simplifications* (18 vs. 43) and no *neutralization* edits, which may be due to data imbalance regarding such types.

In terms of average quality, *specification* (65%) and *disambiguation* edits (63%) most often lead to improvements, but the eight types appear rather balanced in this regard. The Jaccard similarity score between optimizations performed by humans and our approach is 0.37, mostly agreeing on copy edits (178 cases) and corroboration (22 cases). Given such low overlap, future work should consider conditioning models to generate specific optimizations.

7.3 Performance across Revision Domains

Lastly, we examine whether our approach, along with the chosen text quality metrics, applies to texts from other domains. We consider two datasets: *WikiHow* (Anthonio and Roth, 2020), containing revisions of instructional texts, and *IteraTeR* (Du et al., 2022), containing revisions of various formal texts, such as encyclopedia entries, news, and scientific papers. For our experiments, we use the provided document-level splits, and sample 1000

Approach	BLEU	RouL	SARI	NoEd↓	ExM
WikiHow Dataset					
Unedited	65.7	0.85	28.4	1.00	0.00%
BART + Top-1	**64.7**	**0.83**	41.3	0.50	13.0%
BART + AutoScore	61.8	0.80	**48.5**	**0.08**	**16.0%**
IteraTeR Dataset					
Unedited	74.0	0.86	28.6	1.00	0.00%
BART + Top-1	**68.9**	**0.83**	37.0	0.07	0.00%
BART + AutoScore	64.8	0.80	**38.6**	**0.02**	0.00%

Table 6: Automatic evaluation: Performance of candidate selection strategies on data from other domains, in terms of BLEU, Rouge-L, SARI, ratio of unedited samples, and ratio of exact matches to target reference.

revision pairs pseudo-randomly as a final test set.

Table 6 shows automatic evaluation results. In both cases, *BART+Autoscore* leads to higher SARI scores (48.5 vs. 41.3 for WikiHow, 38.6 vs. 37.0 for IteraTeR), and notably reduces the number of cases where the models failed to revise the input (0.08 vs. 0.50 for WikiHow). The reported *BART+Top1* model represents the approach of Du et al. (2022), indicating that our approach and its text quality metrics achieve state-of-the-art performance with systematic improvements across domains, when generating optimized content. However, as different domains of text have different goals, different notions of quality, and, subsequently, different revision types performed, integrating domain-specific quality metrics may further improve performance. We leave this for future work.

8 Conclusion

With this paper, we work towards the next level of computational argument quality research, namely, to not only *assess* but also to *optimize* argumentative text. Applications include suggesting improvements in writing support and automatic phrasing in debating systems. We presented an approach that generates multiple candidate claim optimizations and then selects the best one using various quality metrics. In experiments, combining fine-tuned BART with such candidate selection improved 60% of the claims from online debates, outperforming several baseline models and candidate selection strategies. We showcased generalization capabilities on two out-of-domain datasets, but we also found some claim optimization types hard to automate.

In future work, we seek to examine whether recent large language models (e.g., Alpaca) and end-to-end models (where generation and candidate selection are learned jointly) can further optimize the quality of claims. As our approach so far relies on the availability of large claim revision corpora and language models, techniques for low-resource scenarios and languages should be explored to make claim optimization more widely applicable.

Acknowledgments

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number 374666841, SFB 1342.

Ethical Considerations

This work contributes to the task of argumentative text editing, namely we explore how to revise claims automatically in order to optimize their quality. While our work may also improve downstream task performance on other tasks, it is mainly intended to support humans in scenarios, such as the creation and moderation of content on online debate platforms as well as the improvement of arguments generated or retrieved by other systems. In particular, the presented approach is meant to help users by showing examples of how to further optimize their claims in relation to a certain debate topic, so they can deliver their messages effectively and hone their writing skills.

However, our generation approach still comes with limitations and may favor revision patterns over others in unpredictable ways, both of which might raise ethical concerns. For example, it may occasionally produce false claims based on untrue or non-existent facts. We think, humans should be able to identify such cases in light of the available context though, as long as the improvements remain suggestions and do not happen fully automatically, as intended.

The presented technology might further be subject to intentional misuse. A word processing software, for example, could be conditioned to automatically detect and adapt claims made by the user in subtle ways that favors political or social views of the software provider. Such misuse might then not only change the intended message of the text, but also influence or even change the views of the user (Jakesch et al., 2023).

In a different scenario, online services, such as social media platforms or review portals, might change posted claims (e.g. social media posts, online reviews) to personalize them and increase user

engagement or revenue. These changes might not only negatively affect the posting, but also the visiting user.

While it is hard to prevent such misuse, we think that the described scenarios are fairly unlikely, as such changes tend to be noticed by the online community quickly. Furthermore, the presented architecture and training procedure would require notable adaptations to produce such high-quality revisions.

An aspect that remains unexplored in this work is the ability of the presented approaches to work with variations of the English language, such as African-American English, mainly due to the lack of available data. In this regard, the approach might unfairly disadvantage or favor particular language varieties and dialects, potentially inducing social bias and harm if applied in public scenarios. We encourage researchers and practitioners to stay alert for such cases and to choose training data with care for various social groups.

Finally, our work included the labeling of generated candidate claims on a crowdsourcing platform. As detailed in Section 5, we compensated MTurk workers $13 per hour, complying with minimum wage standards in most countries at the time of conducting the experiment.

References

Tazin Afrin and Diane Litman. 2018. Annotation and classification of sentence-level revision improvement. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–246, New Orleans, Louisiana. Association for Computational Linguistics.

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 224–233, Online. Association for Computational Linguistics.

Talita Anthonio and Michael Roth. 2020. What can we learn from noun substitutions in revision histories? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1359–1370, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. 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)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, and Smaranda Muresan. 2021. ENTRUST: Argument reframing with language models and entailment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4958–4971, Online. Association for Computational Linguistics.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proceedings of COLING 2012*, pages 711–726, Mumbai, India. The COLING 2012 Organizing Committee.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in Wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, Washington, USA. Association for Computational Linguistics.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland. Association for Computational Linguistics.

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

Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2021. Text editing by command. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5274, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–15, New York, NY, USA. Association for Computing Machinery.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 217–226, New York, NY, USA. Association for Computing Machinery.

Iman Jundi, Neele Falk, Eva Maria Vecchi, and Gabriella Lapesa. 2023. Node placement in argument maps: Modeling unidirectional relations in high & low-resource scenarios. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5854–5876, Toronto, Canada. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Ravi Kondadadi, Blake Howald, and Frank Schilder. 2013. A statistical NLG framework for aggregated planning and realization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1406–1415, Sofia, Bulgaria. Association for Computational Linguistics.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

Gabriella Lapesa, Eva Maria Vecchi, Serena Villata, and Henning Wachsmuth. 2023. Mining, assessing, and improving arguments in NLP and the social sciences. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–6, Dubrovnik, Croatia. Association for Computational Linguistics.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3291–3301, Minneapolis, Minnesota. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Santiago Marro, Elena Cabrio, and Serena Villata. 2022. Graph embeddings for argumentation quality assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4154–4164, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aurélien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Tomoya Mizumoto and Yuji Matsumoto. 2016. Discriminative reranking for grammatical error correction with statistical machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1133–1138, San Diego, California. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1718–1729, Online. Association for Computational Linguistics.

Gabriella Skitalinskaya and Henning Wachsmuth. 2023. To revise or not to revise: Learning to detect improvable claims for argumentative writing support. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15799–15816, Toronto, Canada. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models.

Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 340–350, Austin, Texas. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251. Association for Computational Linguistics.

Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 515–522, Toulouse, France. Association for Computational Linguistics.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting cognitive and emotional empathic writing of students. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4063–4077, Online. Association for Computational Linguistics.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying Semantic Edit Intentions from Revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark. Association for Computational Linguistics.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. SOME: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–143, Denver, Colorado. Association for Computational Linguistics.

Yilun Zhou, Julie Shah, and Steven Schockaert. 2019. Learning household task knowledge from WikiHow descriptions. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 50–56, Macau, China. Association for Computational Linguistics.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.

A Implementation and Training Details

A.1 Candidate Generation Models

For generation, we use the pre-trained BART model implemented in the fairseq library. The library and pre-trained models are BSD-licensed. We use the BART-large checkpoint (400M parameters) and further finetune the model for 10 epochs on 2 RTX 2080Ti GPUs. We use the same parameters as suggested in the fine-tuning of BART for the CNN-DM summarization task by fairseq and set MAX-TOKENS to 1024. The training time is 100-140 minutes, depending on the chosen setup (with or without context information).

During inference, we generate candidates using a top-k random sampling scheme (Fan et al., 2018) with the following parameters: length penalty is set to 1.0, n-grams of size 3 can only be repeated once, temperature is set to 0.7, while the minimum and maximum length of the sequence to be generated are 7 and 256 accordingly.

A.2 Quality Assessment Models

For the automatic assessment of fluency and argument quality, we use the bert-base-cased pre-trained BERT version, as implemented in the huggingface library. The library and pre-trained models have the Apache License 2.0. We finetune the model for two epochs and use the parameters suggested in Skitalinskaya et al. (2021). The accuracy of the trained model for fluency obtained on the train/dev/test split suggested by the authors (Toutanova et al., 2016) is 77.4 and 75.5 for argument quality.

For labeling the missing or unassigned revision types, we use the same bert-base-cased pre-trained BERT model, but in a multi-label setup, where we consider the following 6 classes: claim clarification, typo or grammar correction, correcting or adding links, changing the meaning of the claim, splitting the claim, and merging claims. We fine-tune the model for two epochs using the Adam optimizer with a learning rate of 1e-5 and achieve a weighted F1-score of 0.81.

B Alternative Generation Models

For comparison, we provide two additional baseline Seq2Seq model architectures, which help identify the complexity of the model needed for the task:

LSTM. Our first baseline is a popular LSTM variant introduced by Wiseman and Rush (2016). We use the *lstm_wiseman_iwslt_de_e* architecture, which is a two-layer encoder and decoder LSTM, each with 256 hidden units, and dropout with a rate of 0.1 between LSTM layers.

Transformer. The second model is based on the work of Vaswani et al. (2017). We use the *transformer_iwslt_de_en* architecture, a 6-layer encoder and decoder with 512-dimensional embeddings, 1024 for inner-layers, and four self-attention heads.

Tables 7 and 8 compare the automatic evaluation scores of all generation-content selection combinations.

Model	Strategy	BLEU	RouL	SARI	NoEd↓	ExM
BART	Top-1	64.0	0.83	39.7	0.31	7.8%
	Random	62.6	0.83	38.7	0.28	6.8%
	SVMRank	55.7	0.76	38.8	0.03	4.5%
	AutoScore	59.4	0.80	43.7	0.02	8.3%
Trans-former	Top-1	43.6	0.64	0.30	0.12	0.8%
	Random	42.4	0.63	0.30	0.13	1.0%
	SVMRank	41.8	0.63	0.31	0.10	1.2%
	AutoScore	40.5	0.62	0.30	0.10	1.3%
LSTM	Top-1	36.2	0.56	0.28	0.10	0.3%
	Random	36.0	0.56	0.28	0.10	0.3%
	SVMRank	36.2	0.56	0.29	0.10	1.0%
	AutoScore	34.1	0.52	0.28	0.10	1.0%

Table 7: Automatic evaluation: Results for each combination of generation model and candidate selection strategy on the 600 test samples, in comparison to the human revisions: BLEU (0-100), ROUGE-L (RouL), SARI, ratio of unedited samples (NoEd), % of exact matches to target reference (ExM).

Model	Strategy	Fluency	Meaning	Argument	Average
BART	Top-1	0.73	0.97	0.65	0.78
	Random	0.72	0.97	0.68	0.79
	SVMRank	0.72	0.94	0.76	0.81
	AutoScore	0.83	0.95	0.86	**0.88**
Trans-former	Top-1	0.44	0.76	0.40	0.53
	Random	0.41	0.76	0.38	0.52
	SVMRank	0.50	0.76	0.45	0.57
	AutoScore	0.68	0.75	0.61	**0.68**
LSTM	Top-1	0.27	0.68	0.31	0.42
	Random	0.27	0.68	0.31	0.42
	SVMRank	0.29	0.69	0.31	0.43
	AutoScore	0.52	0.65	0.53	**0.57**
Human		0.72	0.94	0.74	0.80

Table 8: Results for each combination of generation model and candidate selection strategy on the 600 test samples, in comparison to the human revisions based on three quality metrics: fluency, meaning preservation and argument quality.

B.1 Automatic Evaluation

We use the following python packages and scripts to perform automatic evaluations: nltk (BLEU (Papineni et al., 2002)), rouge-score (ROUGE (Lin, 2004)), https://github.com/cocoxu/simplification/SARI.py (SARI (Xu et al., 2016))

C Claim Optimization Examples

For all eight optimization categories, we provide one or more examples illustrating each action in Table 9.

D Manual Quality Assessment Guidelines

Figure 3 shows the annotation guidelines for the Amazon Mechanical Turk study.

E System Outputs

Table 10 provides examples of candidates selected by different content selection strategies along with human references illustrating common patterns found in the results. Table 11 provides examples of candidates generated with and without utilizing context knowledge with insertions and deletions being highlighted in green and red fonts accordingly.

Type	Examples
Specification	Nipples are the openings of female-only exocrene glands that can have abnormal [secretions] <LINK> during any time of life, get erected by cold stimulation or sexual excitement (much more visibly than in men), get lumps or bumps and change color and size of areola during the menstrual cycle or pregnancy, so their display can break [personal space] <LINK> and privacy (which is stressful), affect public sensibilities and also be a [window] <LINK> for infections, allergies, and irritation. The idea behind laws, such as limiting the amount of guns, is to reduce the need to defend yourself from a gun or rapist. It is very common for governments to actively make certain forms of healthcare [harder for minority groups to access] <LINK>. They could also, therefore, make cloning technology hard to access.
Simplification	Very complex~~, cognitively meaningful behavior such as~~ behaviours like creating art are evidence of free will~~, because they exhibit the same lack of predictability as stochastic systems, but are intelligible and articulate clearly via recognizable vehicles~~.
Reframing	It reduces the oversight of the BaFin and thus increases ~~the risk of financial crisis~~ market failures.
Elaboration	It takes 2-4 weeks for HIV to present any symptom. The incubation period risk ~~can't be ruled out for~~ is higher for a member of high risk group, ~~effectively and timely~~ even though member of a low risk group is not completely safe. The decision is based on the overall risk, not on individual level.
Corroboration	[Person-based predictive policing technologies] <LINK> - that focus on predicting who is likely to commit crime rather than where is it likely to occur - violate the [presumption of innocence.] <LINK>.
Neutralization	Biden ~~does not~~ lacks the support ~~or agree with several key issues that are important to liberal voters.~~ of many liberal voting groups due to his stance on key issues concerning them.
Disambiguation	The USSR had [passed legislation] <LINK> to gradually eliminate religious belief within its borders. However the death penalty was more used in USSR than in Russia. ~~It~~ USSR had 2000 [death penalties] <LINK> per year in the 1980s whereas pre USSR Russia had [banned the death penalty] <LINK> in 1917 and almost never carried it out in the decades before that. ~~SRM~~ Solar geoengineering merely serves as a "technological fix" (Weinberg).[harvard.edu] <LINK>
Copy Editing	Women are experiencing record ~~level~~ levels of success in primaries.

Table 9: Illustrative examples of optimization types identified in the paper. The green font denotes additions and the striked out red font denotes the removal of text snippets.

Instructions

In this task, your goal is to identify whether a claim has been successfully improved, without changing the overall meaning of the text.
Each task contains a set of pairs, where one claim is the "original claim," and the other an optimized candidate. Each of these pairs have the same original text, but different candidate optimizations.

Please rate each candidate along the following three perspectives: argument quality, fluency and semantic similarity. And, finally, please, rank all candidates relative to each other in terms of overall quality.

Argument Quality

Scale (1-5): 1 (notably worse than original), 2 (slightly worse), 3 (same as original), 4 (slightly improved), 5 (notably improved)

Does the optimized claim improve the argument quality compared to the original claim? Relevant changes include, but are not limited to:

- further specifying or explaining an existing fact or meaning
- removing information or simplifying the sentence structure with the intent to reduce the complexity or breadth of the claim
- rephrasing a claim with the intent to specify or generalize the claim, or to add clarity
- adding (substantive) new content or information to the claim or inserting an additional fact with the intent of making it more self-contained, more sound or stronger
- adding, editing or removing evidence in the form of links that provide supporting information or external resources to the claim
- removal of bias or biased language
- removal uncertainty. e.g. by replacing pronouns referring to concepts that have been mentioned in other claims of the debate, or by replacing acronyms with what they stand for
- improving the grammar, spelling, tone, or punctuation of a claim

Meaning

Scale (1-5): 1 (entirely different), 2 (substantial differences), 3 (moderate differences), 4 (minor differences), 5 (identical)

Does the transformed claim still have the same overall meaning as the original? It is OK if extra information is added, as long as it doesn't change the underlying people, events, and objects described in the sentence. You should also not strongly penalize for meaning transformations which aim to generalize or specify some aspects of the claim.

Fluency

Scale (1-3): 1 (major errors, disfluent), 2 (minor errors), 3 (fluent)

Is this sentence fluent English and does it make sense?

Examples with explanations can be found here.

Figure 3: Guidelines provided to the crowdworkers to annotate automatically generated claim optimizations.

Strategy	Candidate
Original	Easy to cheat, and procastinate
Top-1	It is easy to cheat and procastinate .
SVMRank	The easy to cheat, and procastinate
AutoScore	The UBI is easy to cheat and easy to manipulate.
Human	Flexible schedules can easily increase procrastination
Original	Many games often attempt to depict the norms and attitudes of the real world.
Top-1	Many games often attempt to depict the norms and attitudes of the real world.
SVMRank	Many games often attempt to depict the norms and attitudes of the real world, and thus are likely to be accurate and relevant to what people are thinking and feeling in the real-world.
AutoScore	Many games often attempt to depict the norms and attitudes of the real world, in order to give players a greater understanding of the world around them.
Human	Many games often attempt to depict the norms and attitudes of the real world; gamers are likely to see social attitudes espoused in these games on issues of gender equality as indicative of progress in the real world.
Original	Owners of PWA take home [more profit] <LINK> from their sales than mobile apps.
Top-1	Owners of PWA take home [more profit] <LINK> from their sales than mobile apps.
SVMRank	Owners of PWA take home [more profit] <LINK> from their sales than mobile apps. This allows them to reinvest in their business and increase their profits.
AutoScore	Owners of PWA take home [more profit] <LINK> from their sales than those of mobile apps.
Human	Owners of PWAs take home [more profit] <LINK> from their sales than mobile apps.
Original	Scientists found out that Saturn's moon [Enceladus] <LINK> houses a giant, global ocean beneath its icy outer shell.
Top-1	Scientists found out that Saturn's moon [Enceladus] <LINK> houses a giant, global ocean beneath its icy outer shell.
SVMRank	Scientists found out that Saturn's moon [Enceladus] <LINK> houses a giant, global ocean beneath its icy outer shell. This suggests that this icy, water-filled place also houses other life forms, such as asteroids and comets.
AutoScore	Scientists found out that Saturn's moon [Enceladus] <LINK> houses a giant, global ocean beneath its icy outer shell, suggesting that the moon may hold life beyond the surface.
Human	Scientists found ~~out~~ that Saturn's moon [Enceladus] <LINK> houses a giant, global ocean beneath its icy outer shell.
Original	Responsibility requires conscious decisions and [free will] <LINK>. Acting according to governmental orders is far from that.
Top-1	Responsibility requires conscious decisions and [free will] <LINK>. Acting according to governmental orders is far from that.
SVMRank	Responsibility requires conscious decisions and [free will] <LINK>. Acting according to governmental orders is far from that. Therefore, people are not responsible for their actions, and are therefore not morally responsible.
AutoScore	Responsibility requires conscious decisions and [free will] <LINK>. Acting according to governmental orders is far from that, as it is not a conscious decision but rather a result of an involuntary reaction.
Human	The coercion to vote eliminates the idea of responsibility, which requires conscious decisions and [free will] <LINK>. Acting according to governmental orders is far from that.
Original	There was much progress, advancement and forward looking thinking in the 20th century in Europe across all disciplines.
Top-1	There was much progress, advancement and forward looking thinking in the 20th century in Europe across all disciplines. Therefore, it is logical to conclude that the world has changed in the last 150 years.
SVMRank	There was much progress, advancement and forward looking thinking in the 20th century in Europe across all disciplines, which shows the progress that was made during the century.
AutoScore	There was much progress, advancement, and forward looking thinking in the 20th century in Europe across all disciplines.
Human	There was much progress, advancement and forward ~~looking~~ thinking in the 20th century in Europe across all disciplines.
Original	A [catch-22] <LINK> situation currently exists in regards to researching the medicinal applications of some illegal drugs, due to the laws surrounding how they are categorised ([p. 12] <LINK>).
SVMRank (Top-1)	A [catch-22] <LINK> situation currently exists in regards to researching the medicinal applications of some illegal drugs, due to the laws surrounding how they are categorised ([p. 12] <LINK>). This prevents researchers from knowing how to fully assess the medicinal potential of drugs in a manner that is best suited for the patients who need it.
AutoScore	A [catch-22] <LINK> situation currently exists in regards to researching the medicinal applications of some illegal drugs, due to the laws surrounding how they are categorised ([p. 12] <LINK>). This prevents researchers from researching possible medicinal uses for illegal drugs.
Human	A [catch-22] <LINK> situation currently exists in regard to researching the medicinal applications of some illegal drugs, due to the laws surrounding how they are categorised ([p. 12] <LINK>).

Table 10: Examples of optimizations selected by different candidate selection strategies along with human references. The green font denotes additions and the striked out red font denotes the removal of text snippets.

Control	Candidate
Topic	Should Men Be Involved And Trusted In Childcare?
Previous	Perpetrators of child abuse are [way more likely] <LINK> to be the child's parents.
Original	Mostly the mothers, too.
Claim only	The mothers are the most vulnerable.
+ Previous	The majority of perpetrators are the mothers, too.
+ Topic	It is mostly mothers who are responsible for childcare.
Topic	Would Trump or Biden be a better President?
Previous	The [Healthy Indiana Plan 2.0] <LINK>, initiated by Mike Pence, has provided access to healthcare to many people who could not previously afford it.
Original	The Healthy Indiana Plan has high penalties for non-payment of [premiums] <LINK>.
Claim only	The Healthy Indiana Plan has high penalties for non-payment of [premiums] <LINK>. This means that people with low income could be penalized for failure to pay.
+ Previous	The Healthy Indiana Plan has high penalties for non-payment of [premiums] <LINK> in exchange for providing healthcare.
+ Topic	The Healthy Indiana Plan has high penalties for non-payment of [premiums] <LINK>,which were imposed in 2015.
Topic	Artificial Intelligence (AI): Should an artificial general intelligence be created?
Previous	There is the chance an AGI might be able to think for itself and turn on its creators.
Original	AGI are susceptable.
Claim only	Many AGI are susceptable to the danger of death.
+ Previos	AGI are susceptible. They may be influenced by humans and by the biases of their creators.
+ Topic	AGI are susceptibleto being hacked.
Topic	Who should the Conservative Party choose to be the next UK Prime Minister?
Previous	[Jeremy Hunt] <LINK>
Original	Hunt is the best placed candidate to unify Parliament in order to get a Brexit deal passed.
Claim only	Hunt is the best placed candidate to unify Parliament and the public in order to get a Brexit deal passed.
+ Previous	Hunt could use his position as the Prime Minister to unify Parliament in order to get a Brexit deal passed.
+ Topic	Hunt is the best placed candidate to [unify Parliament] <LINK> in order to get a Brexit deal passed
Topic	Should high-income countries take in refugees?
Previous	Taking in refugees will increase criminality within host countries.
Original	Differences in criminal activity are a product, in part, of childhood [socioeconomic] <LINK> conditions. This is exacerbated by the longer [path] <LINK> to employment faced by refugees compared to other communities.
Claim only	Differences in criminal activity are a product, in part, of childhood [socioeconomic] <LINK> conditions. This is exacerbated by the longer [path] <LINK> to employment faced by refugees compared to other communities, making them more likely to get involved in crime.
+ Previous	Differences in criminal activity are a product, in part, of childhood [socioeconomic] <LINK> conditions. This is exacerbated by the longer [path] <LINK> to employment faced by refugees compared to other communities. This will not increase criminality.
+ Topic	Differences in criminal activity are a product, in part, of childhood [socioeconomic] <LINK> conditions. This is exacerbated by the longer [path] <LINK> to employment faced by refugees compared to other communities, which make it harder to find a job.
Topic	Mark Twain used the N-word in The Adventures of Huckleberry Finn. Should it be censored?
Previous	Changing the N-word would skip a piece of the $linguistic$ past and thus everyday life. As a result, people could start to forget this part of history.
Original	In Huckleberry Finn, Twain captured the essence of "[everyday midwest American English] <LINK>".
Claim only	In Huckleberry Finn, Twain captured the essence of "[everyday midwest American English] <LINK>".This is a common trait of the American English language.
+ Previous	In Huckleberry Finn, Twain captured the essence of "[everyday midwest American English] <LINK>"by using the N-word in everyday conversation.
+ Topic	In Huckleberry Finn, Twain captured the essence of "[everyday midwest American English] <LINK>", which is a language that is often used by people who do not share his values.

Table 11: Examples of different candidates generated by BART + AutoScore with and without context information. The green font denotes additions of text snippets.