Assessing the Sufficiency of Arguments through Conclusion Generation

Timon Gurcke   Milad Alshomary   Henning Wachsmuth
Department of Computer Science
Paderborn University, Paderborn, Germany
`<first name>.<last name>@uni-paderborn.de`

Abstract

The premises of an argument give evidence or other reasons to support a conclusion. However, the amount of support required depends on the generality of a conclusion, the nature of the individual premises, and similar. An argument whose premises make its conclusion rationally worthy to be drawn is called *sufficient* in argument quality research. Previous work tackled sufficiency assessment as a standard text classification problem, not modeling the inherent relation of premises and conclusion. In this paper, we hypothesize that the conclusion of a sufficient argument can be *generated* from its premises. To study this hypothesis, we explore the potential of assessing sufficiency based on the output of large-scale pre-trained language models. Our best model variant achieves an F_1-score of .885, outperforming the previous state-of-the-art and being on par with human experts. While manual evaluation reveals the quality of the generated conclusions, their impact remains low ultimately.

1 Introduction

The quality assessment of natural language argumentation is nowadays studied extensively for various genres and text granularities, from entire news editorials (El Baff et al., 2020) to arguments in online forums (Lauscher et al., 2020) to single claims in social media discussions (Skitalinskaya et al., 2021). The reason lies in its importance for driving downstream applications such as writing support (Stab, 2017), argument search (Wachsmuth et al., 2017b), and debating technologies (Slonim et al., 2021). Wachsmuth et al. (2017a) organized quality dimensions of arguments into three complementary aspects: logic, rhetoric, and dialectic. Logical quality refers to the actual argument structure, that is, how strong an argument is in terms of the support of a claim (the argument's *conclusion*) by evidence and other reasons (the *premises*).

Figure 1: Two example arguments from a persuasive student essay, one classified as *sufficient*, the other as *insufficient* in the corpus of Stab and Gurevych (2017b).

A key dimension of logical quality is *sufficiency*, capturing whether an argument's premises together make it rationally worthy of drawing its conclusion (Johnson and Blair, 2006). Consider, for example, the two arguments on health education in Figure 1, taken from the argument-annotated essay corpus of Stab and Gurevych (2017a). While the upper one was deemed sufficient by human experts (Stab and Gurevych, 2017b), the lower one was not, likely because the second premise tries to reason from a single example. A reliable computational assessment of argument sufficiency would allow systems to determine those arguments that are well-reasoned.

As detailed in Section 2, previous approaches to argument sufficiency assessment model the task as a standard text classification problem and tackle it with convolutional neural networks (Stab and Gurevych, 2017b) or traditional feature engineer-

ing (Wachsmuth and Werner, 2020). In the focused domain of persuasive student essays, Stab and Gurevych (2017b) obtained a macro F_1-score of .827, not far away from human performance in their setting (.887). However, to further improve the state of the art, we expect the integration of knowledge beyond what is directly available in the text at hand is needed. In particular, we observe that existing work neither explicitly considers an argument's premises and conclusions, nor a property of their relationship. We hypothesize that only a sufficient argument makes it possible to infer the conclusion from the premises. Consequently, comparing the stated conclusion of an argument with one that is (automatically) *generated* from the premises could help the model to distinguish sufficient arguments from insufficient ones. This hypothesis raises the question of whether the knowledge encoded in large-scale pre-trained language models can be leveraged, a direction nearly unexplored so far in argument quality assessment.

In this paper, we study whether generating a conclusion from an argument's premises benefits the computational assessment of the argument's sufficiency. In particular, we first enrich the argument with structural annotations, highlighting which parts are the premises and which part is the conclusion. We propose in Section 4 to then mask the conclusion in order to learn to re-generate it using fine-tuned BART (Lewis et al., 2020). Combining the generated conclusion with the original argument and its annotations, our approach learns to distinguish sufficient from insufficient arguments using a modified RoBERTa model (Liu et al., 2019).

Starting from ground-truth argument structure, we subsequently evaluate conclusion generation and sufficiency assessment on the merged annotations of the corpora of Stab and Gurevych (2017a) and Stab and Gurevych (2017b), as described in Section 3. Our generation experiments indicate that fine-tuning BART leads to better conclusions, which are on par with human-written conclusions in terms of sufficiency, likeliness, and novelty (Section 5). To quantify the impact on sufficiency assessment, we explore various combinations of premises, original conclusion, and generated conclusion in systematic ablation tests, and we compare them to the state of the art and a human upper bound (Section 6). Our sufficiency experiments reveal that, even on the plain input text of an argument, RoBERTa already improves significantly

over the state of the art. The addition of structural annotations and the generated conclusion lead to further improvements, although the benefit of generation ultimately remains limited, possibly due to the generally limited importance of knowing the conclusion on the given data. Finally, we discuss the results of our approaches in Section 7 in light of their implications for the field.

The main contributions of this paper are:[1]

- A language model that can generate human-like argument conclusions.

- The new state-of-the-art approach to argument sufficiency assessment.

- Insights into the importance of mined and generated structure within argument assessment.

2 Related Work

Computational argumentation research has assessed various dimensions of argument quality. Wachsmuth et al. (2017a) provide a theory-based taxonomy of 15 logical, rhetorical, and dialectical quality dimensions and of the work in natural language processing done in these directions. We focus on the (local) *sufficiency* dimension, which is key to logical cogency, representing that an argument's conclusion can rationally be drawn from its premises, given that these are acceptable and relevant (Johnson and Blair, 2006).

Few approaches tackled sufficiency computationally so far. Aside from Wachsmuth and Werner (2020) who assess it as one of the 15 dimensions above using traditional text-focused feature engineering, we are only aware of the work of Stab and Gurevych (2017b) who extend the argument-annotated essay corpus of Stab and Gurevych (2017a) with binary sufficiency annotations. On this basis, the authors compare a support vector machine using lexical, syntactic, and length features to a convolutional neural network (CNN) with word vectors, the latter achieving the best result with a macro F_1-score of .827. In our experiments, we use their dataset and replicate their experiment settings, in order to compare to the CNN.

Unlike Stab and Gurevych (2017a), we rely on a transformer-based architecture, namely we adapt RoBERTa (Liu et al., 2019) to assess sufficiency. Approaches to argument quality assessment using such architectures are still limited, mostly focusing

---

[1]The experiment code can be found under: https://github.com/webis-de/ArgMining-21

on a holistic view of quality (Gretz et al., 2020; Toledo et al., 2019), although a few approaches used transformers for some of the dimensions of Wachsmuth et al. (2017a), such as Lauscher et al. (2020), or somewhat related dimensions in light of quality improvement (Skitalinskaya et al., 2021).

In contrast to the standard use of transformers for text classification, we leverage the structure of arguments for their assessment. Wachsmuth et al. (2016) provided evidence that mining the argumentative structure of persuasive essays helps to better assess four essay-level quality dimensions of argumentation. Similarly, we use annotations of the premises and conclusions of arguments for the sufficiency assessment, but we target the arguments. Moreover, we explore to benefit of conclusion *generation* for the assessment.

The idea of reconstructing an argument's conclusion from its premises was introduced by Alshomary et al. (2020), but their approach focused on the inference of a conclusion's target. The actual generation of entire conclusions has so far only been studied by Syed et al. (2021). The authors presented the first corpus for this task along with experiments where they adapted BART (Lewis et al., 2020) from summarization to conclusion generation. While they trained BART to directly generate a conclusion based on premises, we generate conclusions that fit the context of an entire argument. To this end, we leverage and finetune BART's inherent denoising capabilities obtained during pretraining to replace a mask token in an argument.

3 Data

To study our hypothesis that a sufficient argument's conclusion can be generated from its premises, we need data that is annotated for both argument structure and sufficiency. In this section, we describe how we employ existing corpora for this purpose.

3.1 Data for Conclusion Generation

The argument-annotated essay (AAE-v2) corpus (Stab and Gurevych, 2017a) contains structural annotations for 402 complete persuasive student essays. For conclusion generation (as well as for structure-based sufficiency assessment), we only need annotations of single arguments.

Our instance creation procedure resembles the one of Alshomary et al. (2020), but we work on argument level rather than conclusion level, since we approach conclusion generation as a language

model denoising task (Lewis et al., 2020). Concretely, instead of using premises-conclusion training pairs where the conclusion shall be generated given the premises, we rely on argument-argument pairs. The first argument here is a modified version of the second argument where the conclusion is masked. This way, we avoid conflicts with the argument-level sufficiency annotations of Stab and Gurevych (2017b) (see below). In total, we obtain 1506 argument-argument pairs relating to 1506 unique conclusions matched with 1029 unique arguments. On average, each argument has a length of 4.5 sentences and contains 94.6 tokens.

For training, we rely on 5-fold cross-validation, ensuring that the argument-argument pairs from one essay are never split between training, validation, and test data. This prevents possible data leakage that could artificially improve the final evaluation scores. For each folding, we use 70% training, 10% validation, and 20% test data.

3.2 Data for Sufficiency Assessment

Stab and Gurevych (2017b) further classified each argument in the 402 essays of the AAE-v2 corpus as being sufficient or not. Following Johnson and Blair (2006), the authors defined that an "argument complies with the sufficiency criterion if its premises provide enough evidence for accepting or rejecting the claim" (we speak of "conclusion" here instead of "claim"). All 1029 arguments were labeled, of which 681 (66.2%) were considered sufficient and 348 (33.8%) insufficient.

We use the provided corpus both in its original form and in a modified version where we replace the conclusion of an argument with two separator tokens, "</s></s>". This allows us to study a wide range of different approaches for sufficiency assessment by placing text in-between the two tokens as a replacement for the original conclusion.

For training, we replicate the original 20-times 5-fold cross-validation setup of Stab and Gurevych (2017b), with 70% training, 10% validation, and 20% test data, in order to ensure comparability.

4 Approach

This section describes our two-step approach to assess the sufficiency of a given argument through conclusion generation. First, we generate a conclusion from the argument's premises using a pretrained language model finetuned on the task of replacing the masked conclusion of an argument.
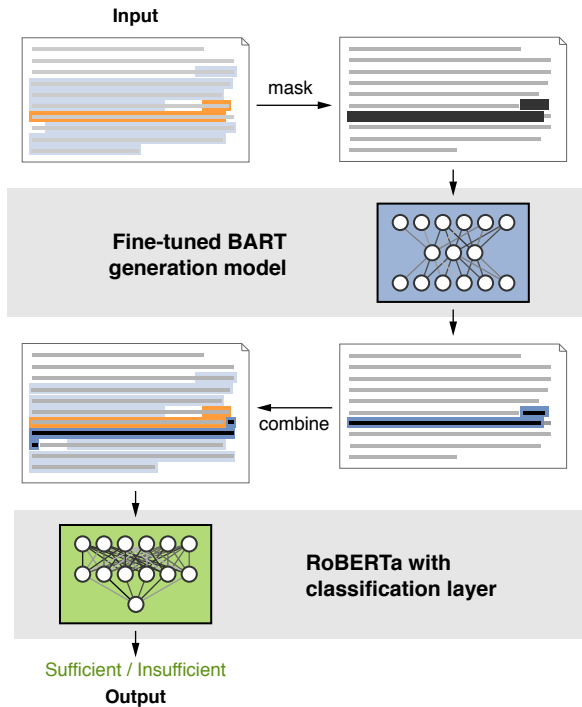
Figure 2: Illustration of our sufficiency assessment approach through generation: (1) BART is used to generate the masked conclusion in an argument. (2) The generated conclusion is combined with the ground truth annotation of the argument. (3) RoBERTa classifies the enriched argument as sufficient/insufficient. Several ablations of the annotations are tested in our experiments.

Second, the generated conclusion is used to assess the argument's sufficiency by experimenting with eight modified versions of the original input argument (Section 6). An overview of the approach is shown in Figure 2. In the following, we detail how we train the models for the two steps.

4.1 Conclusion Generation using Denoising

Given an argument with a masked conclusion, the first task is to re-generate the conclusion. To tackle this task, we use BART-large (Lewis et al., 2020) and treat generation as a denoising task. We explore two model variants:

BART-unsupervised In this variant, we do not finetune BART on any data, but we use its vanilla denoising capabilities obtained in its pre-training procedure. Note that the masked conclusions usually do not represent entire sentences, thus leaving textual markers which trigger BART to generate a logical conclusion, for example, "Thus, <mask>" or "This makes it clear that <mask>." We consider this model as a baseline.

BART-supervised In this variant, we finetune BART on the data from Section 3, in order to tailor its denoising capabilities towards conclusion generation. In particular, we thereby adjust the language model towards the given domain and teach the model to replace the mask token with a conclusion (instead of just generating text that fits the context). We finetune BART using cross-entropy loss, as commonly done in text generation.

The proper training settings of the two models are found via a hyperparameter search. For the evaluation described below, we ran 10 trials for each fold, testing batch sizes between 4 and 8 and learning rates between $5 \cdot 10^{-6}$ and $5 \cdot 10^{-5}$. We fixed the number of epochs to 3 per fold, as we did not observe any improvements afterwards, and we used a cosine learning rate scheduler with 50 warmup steps to stabilize the training. At inference time, we employed a beam size of 5 to obtain the final generated conclusions. We considered the epoch out of three, which performs best on the validation data in terms of BERTScore (Zhang* et al., 2020).

4.2 Sufficiency Assessment using Structure

Given a modified argument, the second task is to predict whether the premises in the argument are rationally worth drawing the conclusion. We use RoBERTa (Liu et al., 2019) for this task by adding a linear layer on top of the pooled output of the original model (Devlin et al., 2019).

Successfully optimizing RoBERTa using mean squared error (MSE) or cross-entropy loss functions would be difficult, as both of them do not align well with the target metric of sufficiency assessment (macro F_1-score). We therefore follow ideas of Puthiya Parambath et al. (2014) and Eban et al. (2017) who propose to optimize machine learning models on the F_1-score directly. Accordingly, we allow the model to output probabilities instead of interpreting a single binary value. Analogous to Stab and Gurevych (2017b), we allow our model to adjust hyperparameters between folds. In our experiments, we followed the same hyperparameter optimization procedure as before but for different parameters and ranges. In total, we ran 10 trials for each fold, and we adjusted the batch size to be between 16 and 32 and the learning rate between 10^{-6} and $5 \cdot 10^{-5}$. We selected the epoch for each trial out of three, which performed best on the validation data in terms of macro F_1-score.

70

Model	BERTScore	ROU.-1	ROU.-2	ROU.-L
BART-unsupervised	0.14	19.69	4.05	16.40
BART-supervised	**0.25**	**20.97**	**4.79**	**17.49**

Table 1: Automatic evaluation results of concluion generation: Rescaled F1-BERTScore and ROUGE-1/-2/-L scores of the two considered models on the full corpus.

5 Evaluation of Conclusion Generation

To study our hypothesis, we need to ensure that the generated conclusions are meaningful and fit in the context of a given argument, so they can be helpful in sufficiency assessment. In this section, we therefore evaluate the quality of the generated conclusions, both automatically and manually.

5.1 Automatic Evaluation

As indicated in Section 4, we compare two approaches: (1) *BART-unsupervised*, which replaces the mask token in an argument (denoising) with fitting text, as it is part of BART's training procedure (Lewis et al., 2020); and (2) *BART-supervised*, which finetunes BART on the argument-argument pairs from Section 3. For both approaches, we obtained the complete set of generated conclusions using the cross-validation setup described in Section 4. Matching these with the corresponding ground-truth conclusion, we then computed their quality in terms of BERTScore as well as ROUGE-1, ROUGE-2, and ROUGE-L.

Results   Table 1 lists the results of the approaches. BART-unsupervised is a strong baseline in terms of lexical accuracy: Values such as 19.69 (ROUGE-1) and 16.40 (ROUGE-L) are comparable to those that Syed et al. (2021) achieved in similar domains with sophisticated approaches. However, finetuning on the argument-argument pairs does not only significantly increase the semantic similarity between generated and ground truth conclusions from 0.14 to 0.25 in terms of BERTScore, but it also leads to a slight increase in lexical accuracy.

5.2 Manual Evaluation

The metrics used to automatically evaluate conclusion generation are not ideal, since they expect a single correct result. The task of conclusion generation, in contrast, allows for multiple, possibly very different correct conclusions, for example, a different conclusion target may be derived from a single set of premises (Alshomary et al., 2020). We thus conducted an additional manual annotation study

to evaluate the quality of the conclusions generated by the two approaches in comparison to the human ground truth.

We randomly chose 100 arguments from the given corpus, 50 labeled as sufficient and 50 labeled as insufficient. For each arguments, we additionally created two variants, replacing the original conclusion with the generated conclusion of either approach. In each case, we then presented the three arguments with their premises and conclusions highlighted to five annotators of different academic backgrounds (economics, computer science, health/medicine), none being an author of this paper. We asked each annotator three questions, Q1–Q3, on each argument, resulting 300 annotations for each model and 900 annotations in total. For consistency reasons, we used a 5-point Likert scale for each question:

- Q1: Are the premises sufficient to draw the conclusion? This question referred to the *sufficiency* of arguments, from "not sufficient" (score 1) to "sufficient" (score 5). We asked this question to see how the sufficiency of generated and human-written conclusions differs, thus directly evaluating our hypothesis.

- Q2: How likely is it that the conclusion will be inferred from the context? This question referred to the *likelihood* of a conclusion, from "very unlikely" (score 1) to " very likely" (score 5). We asked this question as an internal quality assurance, ruling out the possibility that our models generate conclusions unrelated to the given context of the argument.

- Q3: How can the conclusion be composed from the context? This question, finally, referred to the *novelty* of the generated conclusions in light of their context. The score range here is more complex; inspired by Syed et al. (2021), who also ask annotators about the novelty of generated conclusions, it reflects the cognitive load required to infer the conclusions from the context of the argument: "verbatim copying" (1), "synonymous copying" (2), "copying + fusion" (3), "inference" (4), and "can not be composed" (5).

Results   For each question, Table 2 shows the inter-annotator agreement, the distribution of majority scores, and the resulting mean score of the three compared approaches (including the ground truth), and the mean rank. We obtained the rank by

#	Question	Approach	(a) Agreement		(b) Majority Scores					(c) Mean Results	
			α	Majority	1	2	3	4	5	Score ↑	Rank ↓
Q1	Are the premises sufficient to draw the conclusion?	Ground Truth	.23	**63%**	7	24	39	30	0	**2.92**	1.43
		BART-unsupervised	**.43**	**63%**	14	19	34	32	1	2.87	1.40
		BART-supervised	.35	60%	14	21	30	34	1	2.87	**1.37**
Q2	How likely is it that the conclusion will be inferred from the context?	Ground Truth	.19	57%	4	18	54	24	0	**2.98**	1.42
		BART-unsupervised	**.32**	**64%**	16	15	46	23	0	2.76	1.54
		BART-supervised	.28	**64%**	12	15	38	35	0	2.96	**1.39**
Q3	How can the conclusion be composed from the context?	Ground Truth	.19	72%	0	4	18	73	5	**3.79**	**1.39**
		BART-unsupervised	**.50**	80%	14	11	15	47	13	3.34	1.54
		BART-supervised	.32	**82%**	8	8	22	53	9	3.47	1.54

Table 2: Manual evaluation results of conclusion generation on the 100 arguments of each of the three approaches: (a) Agreement of all five annotators in terms of Krippendorff's α and majority. (b) Distribution of majority scores. For Q1/Q2, higher scores mean more sufficient/likely. For Q3, they mean less "copying" (see text for details). (c) Mean score of each approach and rank obtained by comparing the majority score for each argument in isolation.

treating each question as a ranking task where, for each argument, the approaches are ranked from 1 to 3 by decreasing highest majority score.

We find that the general *agreement* for the questions in terms of Krippendorff's α is low, with values between .19 and .50, but comparable to other tasks in the realm of argumentation (Wachsmuth et al., 2017a). On all three questions, the annotators agreed mostly for BART-unsupervised, followed by the BART-supervised, while having the least agreement for the ground truth. This may indicate a more apparent connection of the generated conclusions to the premises. In 57% to 64% of the cases, we observe majority agreement of the annotators for the first two questions, whereas this value goes up to 72%–78% for the last question.

The ranking for Q1 shows that conclusions generated by our BART-supervised model overall ranked best in *sufficiency* (mean rank 1.37), even though the mean score is slightly better for the ground truth. While the differences are small, this behavior is expected as half of the provided arguments were initially labeled as insufficient.

Considering the *likelihood* of the premises (Q2), we see that conclusions of BART-supervised are on par with the ground truth conclusions (score rank 2.96 vs. 2.98, rank 1.39 vs. 1.42) and better than those generated by the baseline BART-unsupervised (1.54). This suggests that the conclusions generated by our model both fit the context of the argument and are at least as likely to be drawn as the ones written by humans. This property is essential for our approach to sufficiency assessment to rule out the possibility of failure due to a low quality of the generated conclusions in general.

Finally, we consider the cognitive load that is required to compose a conclusion given its context, as reflected by *novelty* (Q3). This information is vital to rule out that the generated conclusions are copied from the context of an argument instead of being inferred. We find the ground-truth conclusions to require the most cognitive load in this regard, having a clearly better mean rank (1.39) than the others (both 1.54). Thus, they potentially provide the most novelty to the context. The mean scores indicate an increase of novelty from BART-unsupervised (3.34) to BART-supervised (3.47) though.

The scores and ranks can be interpreted more easily when looking at the *majority scores* for the two BART models and the ground truth. As expected, we find that the amount of conclusions considered to be sufficient is approximately the same to those considered insufficient. The annotators did not agree on the highest sufficiency rank, which may be due to subjectivity in the perception of "full" sufficiency. We observe an analog behavior for score 5 for the likelihood of conclusions (Q2). Here, this may imply that, rarely, only a single conclusion would fit the premises of an argument. Regarding the novelty of the generated conclusions, Q3 reveals that the ground truth annotations are mostly inferences (score 4) and only rarely copied from the context (scores 1 and 2). While BART-unsupervised only somewhat follows this distribution, our BART-supervised model shows a similar behavior to the ground truth, though in less clear form. This is another indication that BART-supervised has learned to generate conclusions from premises.

#	Label	Part	Text
(a) Sufficient		Argument	Second, `<MASK>`. Averagely, public transports use much less gasoline to carry people than private cars. It means that by using public transports, the less gas exhaust is pumped to the air and people will no longer have to bear the stuffy situation on the roads, which is always full of fumes.
		Ground truth	public transportation helps to solve the air pollution problems
		BART-unsupervised	public transport is more efficient than private cars
		BART-supervised	using public transports will help to reduce the amount of pollution in the air
(b) Insufficient		Argument	Last, `<MASK>`. Playing musical instrument is a good way, I can play classical guitar. When I meet difficulties in studies, I will take my guitar and play the song Green Sleeves. It makes me feel better and gives me the confidence.
		Ground truth	we should develop at least one personal hobby, not to show off, but express our emotion when we feel depressed or pressured
		BART-unsupervised	but not least, I love music
		BART-supervised	playing musical instrument is very important to me
(c) Insufficient		Argument	In addition to this, `<MASK>`. For instance, further enforcement banned smoking in capital in Sri Lanka has reduced this consumption related diseases and deaths, as per the ministry of health. As this shows, smoking restrictions has successfully daunted public from this bad puffing that put less strain on country's healthcare systems.
		Ground truth	introducing smoking ban in public places would greatly discourage people from engaging tobacco puffing
		BART-unsupervised	Sri Lankan government has taken several measures to curb smoking
		BART-supervised	smoking restrictions in Sri Lanka has brought a lot of benefits to the country

Table 3: Conclusions generated by the BART models for four arguments (with masked conclusion) compared to the *ground truth* conclusion: (a) BART-supervised almost reconstructs the ground truth. (b) Here, the two models increase sufficiency. (c) Sometimes, the generated conclusions remain rather vague and pick the wrong target.

Error Analysis & Examples To better understand the differences between the two BART models and the ground truth, we analyzed the 100 examples from our annotation study manually. Table 3 compares the conclusions for three arguments.

For the sufficient argument in Table 3(a), BART-supervised nearly perfectly reconstructs the ground truth, whereas BART-unsupervised generates a reasonable but less specific alternative. In Table 3(b), the approaches appear to make the argument more sufficient, particularly BART-supervised. Both examples speak for the truth of our hypothesis that the conclusion of sufficient arguments can be generated from the premises. This tendency is further backed by our analysis in which we found that for BART-supervised 20% (10/50) of the sufficient arguments have perfect matching conclusions (Table 3(a)) 60% (30/50) either are less abstract or have a different conclusion target but are of equal quality, and 20% (10/50) have a conclusion that is of lower quality, compared to the ground truth. In contrast, only 10% (5/50) of the insufficient arguments are perfect matches, 60% (30/50) either are less abstract or have a different conclusion target of

equal quality, and 30% (15/50) have lower quality generated conclusions (Table 3(b)).

Further analysis reveals that for 58% (29/50) of the insufficient arguments, BART-supervised generated a sufficient conclusion with a different target (16/29) or a different level of abstraction, either more specific (11/29) or more abstract (2/29) than the ground truth. In Table 3(c), the BART models seem tricked by the anecdotal evidence given in the premises, mistakenly picking Sri Lanka as the conclusion's target.

Regarding the difference of our models, we found that BART-supervised more often generated conclusions with targets that are equally likely to the target of the human ground truth (68/100 vs. 42/100).

6 Evaluation of Sufficiency Assessment

Using our conclusion generation model, BART-supervised, we finally study our hypothesis on sufficiency assessment by experimenting with different input variations and testing on the corpus of Stab and Gurevych (2017b). The first setting is by just using the RoBERTa model (*direct sufficiency as-*

Assessment	Approach	Accuracy	Macro Pre.	Macro Rec.	Macro F1
Direct	Human upper bound	.911 ± .022	.873 ± .042	.903 ± .020	.883 ± .029
	CNN (Stab and Gurevych, 2017a)	.846 ± .022	.830 ± .021	.832 ± .028	.831 ± .023
	RoBERTa	**.889** ± .026	**.882** ± .057	**.880** ± .078	**.876** ± .031†
Indirect	RoBERTa-premises-only	.887 ± .031	.876 ± .051	.881 ± .071	.875 ± .037
	RoBERTa-conlusion-only	.641 ± .036	.582 ± .048	.567 ± .144	.553 ± .063
	RoBERTa-generated-only	.632 ± .025	.560 ± .038	.544 ± .106	.532 ± .043
	RoBERTa-premises+conclusion	**.896** ± .025	**.888** ± .061	.887 ± .048	**.885** ± .029†‡
	RoBERTa-premises+generated	.889 ± .028	.879 ± .052	.883 ± .064	.878 ± .030
	RoBERTa-conclusion+generated	.659 ± .026	.762 ± .030	.396 ± .092	.571 ± .036
	RoBERTa-all	**.896** ± .024	.886 ± .045	**.889** ± .054	**.885** ± .025†‡

Table 4: Results of argument sufficiency assessment: Accuracy as well as macro precision, recall, and F_1-score of all evaluated approaches, averaged over twenty 5-fold cross-validations. Significant gains over Stab and Gurevych (2017a) and the RoBERTa approach without structural enrichment are marked with † and ‡, respectively (computed using Wilcoxon signed-rank test, p-value .05). The human upper bound is obtained on a subset of 432 arguments.

sessment) and the second by introducing structural knowledge and our generated conclusions (*indirect* sufficiency assessment).

6.1 Direct Sufficiency Assessment

First, we compare the "base version" of our approach, RoBERTa without additional structure annotations, to the *human upper bound* and the state of the art *CNN* of Stab and Gurevych (2017b).[2]

Results The upper part of Table 4 shows the direct assessment results. Our RoBERTa model significantly outperforms the CNN both on accuracy (.889 vs. .846) and on macro F_1 score (.876 vs. .831), the latter being an improvement of whole 4.5 points. Our model also performs almost on par with the human upper bound, meaning it is approximately at the level of human performance. This underlines the potential of pre-trained transformer models in argument quality assessment.

6.2 Indirect Sufficiency Assessment

As our conclusion generation model starts from structural annotations of a given argument, we systematically study the benefit of knowing the premises and the original conclusion, as well of having the generated conclusion. We consider the following eight input variants for the assessment:

- *RoBERTa-premises-only.* Use the full argument as input, but replace the ground-truth conclusion with an <\unk> token.

- *RoBERTa-conclusion-only.* Use only the ground-truth conclusion as input.

- *RoBERTa-generated-only.* Use only the generated conclusion as input.

- *RoBERTa-premises+conclusion.* Use the full argument as input and highlight the ground-truth conclusion using <\s> tokens.

- *RoBERTa-premises+generated.* Use the full argument as input, but replace the ground-truth conclusion with its generated counterpart. Highlight the latter using <\s> tokens.

- *RoBERTa-conclusion+generated.* Use only the ground-truth and the generated conclusion as input, separated with a <\s> token

- *RoBERTa-all.* Use the full argument as input, insert the highlight the generated conclusion after the ground-truth conclusion and highlight both together using <\s> tokens.

Results The lower part of Table 4 shows that both *RoBERTa-premises+conclusion* and *RoBERTa-all* yield the best results overall, significantly outperforming our vanilla *RoBERTa* model by almost 1 point in terms of both accuracy (.896 vs. .889) and macro F_1-score (.885 vs. .876). The results suggest that using a generated conclusion for the argument does not really help, but also not hurts the model performance. This is also supported by *RoBERTa-premises-only*, which also matches the performance of *RoBERTa-premises+generated*. In general, however, bringing structural knowledge to the model gives a slight but significant improvement in sufficiency assessment.

---

[2]Note that the human upper bound of Stab and Gurevych (2017b) was computed on a subset of 433 arguments annotated by three annotators only. Thus, it is only an approximation of the actual human performance. The human scores are based on pairwise comparisons of the three annotators.

Looking at the weak performance of *RoBERTa-conclusion+generated* (macro F$_1$-score .571), we see that an opposition of the two conclusions alone is not enough four sufficiency assessment. Even though adding the generated one improves over having the ground-truth conclusion only, all variants that include the premises perform much better.

To better understand the role of premises and conclusions in sufficiency assessment, we trained the three *RoBERTa-<xy>-only* variants. The high performance of *RoBERTa-premises-only* (macro F$_1$-score .875) clearly reveals that the conclusion is of almost no importance on the data of Stab and Gurevych (2017b), being not significantly worse than vanilla RoBERTa. The low scores of *RoBERTa-conclusion-only* further support this hypothesis, suggesting that the knowledge obtained from the conclusion can be inferred from the argument without its conclusion alone. This result is very insightful in that it displays that the currently available data barely enables a study of sufficiency assessment in terms of its actual definition. Instead, we suppose that models mainly learn a correlation between the nature and the quality of a given set of premises and the possible sufficiency evolving from this. In particular, students who provide "good" premises for an argument in their essays can also deliver an inferrable conclusion.

7 Discussion

Our results suggest that large-scale pre-trained transformer models can help assess the quality of arguments, here their sufficiency. They even nearly matched human performance. However, an accurate understanding of the argument sufficiency task in terms of the actual definition of sufficiency seems barely possible on the available data.

Employing knowledge about argumentative structure can benefit sufficiency assessment, in line with findings on predicting essay-level argument quality (Wachsmuth et al., 2016). Our results suggest that there is at least some additional knowledge in an argument's conclusion that our model could not learn itself. However, we did not actually *mine* argumentative structure here, but we resorted to the human-annotated ground truth, which is usually not available in a real-world setting. Thus, the improvements obtained by the structure could vanish as soon as we resort to computational methods. We note, though, that we obtained state-of-the-art results also using RoBERTa on the plain text only.

Regarding the central hypothesis of this work, we study an example of a more task-aligned approach. However, our results show that answering the question of conclusion inferability by generating conclusions for an argument is difficult, as generated conclusions may be perceived as sufficient, likely, and novel, but may still not be unique. That is, for many premises, it may be possible to generated multiple sufficient conclusions, which naturally limits the impact of *quality assessment through generation*.

Consequently, although the task of sufficiency assessment appears to be solved on the given data, we argue for the need for more sophisticated corpora that better reflect the actual definition of the task, to ultimately allow studying whether approaches such as ours are needed in real-world applications.

8 Conclusion

In this work we have studied the task of argument sufficiency assessment based on auto-generated argument conclusions. According to our findings, traditional approaches can be improved by using large-scale pre-trained transformer models and by incorporating knowledge about argumentative structure. The effect of our proposed idea to leverage generation for the assessment turned out low though. However, this may likely be caused by the available data, wehere sufficiency seems to barely depend on the arguments' conclusions, thus preventing our and previous approaches from actually tackling the task as intended by its definition.

In general, the insights of this paper lay the foundation for more task-oriented approaches towards the assessment of argument quality dimensions, that are tailored towards the properties in scope (here the relation between premises and conclusion). To adequately evaluate such approaches, also refined corpora may be needed.

References

Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. 2020. Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elad Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Ryan Rifkin, and Gal Elidan. 2017. Scalable learning of non-decomposable objectives. In *Artificial intelligence and statistics*, pages 832–840. PMLR.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Ralph Henry Johnson and J Anthony Blair. 2006. *Logical self-defense*. Idea.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. 2014. Optimizing f-measures by cost-sensitive classification. *Advances in Neural Information Processing Systems*, 27:2123–2131.

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.

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.

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Matthias Edwin Stab. 2017. *Argumentative writing support by means of natural language processing*. Ph.D. thesis, Technische Universität Darmstadt.

Christian Matthias Edwin Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990.

Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating informative conclusions for argumentative texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3482–3493, Online. Association for Computational Linguistics.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

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.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. 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.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.