Unsupervised argument reframing with a counterfactual-based approach

Philipp Heinisch
Bielefeld University
pheinisch@techfak.uni-bielefeld.de

Dimitry Mindlin
Bielefeld University
dimitry.mindlin@uni-bielefeld.de

Philipp Cimiano
Bielefeld University
cimiano@techfak.uni-bielefeld.de

Abstract

Framing is an important mechanism in argumentation, as participants in a debate tend to emphasize those aspects or dimensions of the issue under debate that support their standpoint. The task of reframing an argument, that is changing the underlying framing, has received increasing attention recently. We propose a novel unsupervised approach to argument reframing that takes inspiration from counterfactual explanation generation approaches in the field of eXplainable AI (XAI). We formalize the task as a mask-and-replace approach in which an LLM is tasked to replace masked tokens associated with a set of frames to be eliminated by other tokens related to a set of target frames to be added. Our method relies on two key mechanisms: framed decoding and reranking based on a number of metrics similar to those used in XAI to search for a suitable counterfactual. We evaluate our approach on three topics using the dataset by Ruckdeschel and Wiedemann (2022). We show that our two key mechanisms outperform an unguided LLM as a baseline by increasing the ratio of successfully reframed arguments by almost an order of magnitude.

1 Introduction

Framing is an important mechanism in argumentation, as participants in a debate tend to emphasize those aspects or dimensions of the topic under debate that support their standpoint (Misra et al., 2016; Mou et al., 2022). In this context, reframing is a task that has recently received increased attention, consisting in switching the underlying framing of an argument (Chakrabarty et al., 2021; Chen et al., 2021).

In our conceptualization of the problem, there are frames to be deleted, D, and frames to be added, A, to an argument. Our approach essentially masks the tokens that belong to frame D and uses a language model to regenerate the tokens so that ideally they belong to A. Instead of rewriting complete sentences as in previous work (Chen et al., 2021), our approach aims to maximize the change in framing by minimal precise and controlled intervention into the argument. This "mask-and-replace" approach circumvents the need to fine-tune a language model for the specific task and is thus unsupervised.

Consider the following argument debating "nuclear energy" that emphasizes aspects related to *safety*: *"While geothermal, solar, and wind are safe, nuclear energy is not"*. A minimal change to the argument that changes the frame from focusing on safety aspects towards emphasizing economic aspects could yield the following argument: *"While geothermal, solar, and wind are affordable, nuclear energy is not"*.

In this paper, we draw inspiration from current eXplainable AI (XAI) approaches to propose a novel reframing approach that is based on counterfactual explanation generation to explain the decision of a classifier (Wachter et al., 2017). A counterfactual can plainly speaking be seen as an answer to the question: *How would an example have to be different to belong to a different class?* We transfer this idea to the task of reframing arguments, coming up with a "counterfactual" that answers the question: *How would the argument need to be changed to have a different frame?* Counterfactual generation can be seen as a search in the space of possible changes to a given example or argument that switches the class or respective frame. Different metrics have been proposed to constrain and guide the search in the space of possible counterfactuals. As two examples, approaches have used the following metrics: *proximity*, which measures the similarity of the generated instance to the initial instance, and *data manifold closeness*, which measures how well the generated counterfactual fits within the target data distribution (Verma et al., 2020).

Our approach in particular works on the token

Proceedings of the 10th Workshop on Argument Mining, pages 107–119

level, assuming that each frame-relevant token of an argument is assigned to a frame class. The task of frame classification on the token level of an argument has been proposed by Ruckdeschel and Wiedemann (2022). Given the feasibility of this task, we build on this representation and use the models by Ruckdeschel and Wiedemann (2022) as a starting point for our reframing approach.

Our contributions are:

- We present an unsupervised approach to argument reframing that relies on a mask-and-replace approach on the token level, relying on a language model to replace tokens associated with a set of frames to be deleted by other tokens denoting a set of target frames to be added.
- The approach in particular relies on a frame-guided decoding and reranking strategy inspired by the metrics used in counterfactual generation. Concerning the reranking strategy, we transfer existing metrics used in counterfactual explanation generation and adapt them for the case of the reframing task.
- We conduct a comprehensive analysis and evaluation on three controversial topics (nuclear energy, minimum wage, and marijuana), demonstrating the impact of our reranking and framed decoding strategies. We show in particular that these two mechanisms are effective, increasing the ratio of appropriately reframed arguments from 2% to 18% compared to a baseline in our manual evaluation, corresponding to an improvement of almost an order of magnitude. In addition, we analyze the influence of the number of generated candidates as well as of LLM size.

The manual annotations, spanning over 600 reframed arguments as well as our code are available on GitHub[1].

2 Related work

The automatic analysis of frames in texts has been pioneered by Boydstun et al. (2014) and Card et al. (2015), who applied it to the analysis of newspaper articles. Frames help to organize and structure text and arguments but are also used to bias discussions (Mou et al., 2022) or tailor arguments to specific audiences (de Vreese, 2005; Ajjour et al., 2019; Chen et al., 2021).

The task of reframing arguments, as we consider, has been tackled before by Chen et al. (2021), who used the generic frame classes defined by Card et al. (2015) and relied on fine-tuned language models to rewrite complete sentences, using two surrounding sentences as context.

Similar to our goal of minimal changes, Chakrabarty et al. (2021) extended this approach to generate a reframed argument that is closely related the original one. They propose an approach that first identifies parts of the original argument to be replaced and then relies on a fine-tuned BART model to generate replacement candidates, picking the candidate that has the highest score of being entailed by the original argument according to an entailment model.

In contrast to the above-mentioned previous work on reframing that relies on models fine-tuned for the task, our approach is unsupervised.

Beyond the inventory of 15 generic frames proposed by Boydstun et al. (2014), recent work has made a strong case for more fine-granular and topic-specific framesets. Ajjour et al. (2019) have for example explored an approach by which frame labels can be derived bottom-up by clustering, and Mou et al. (2022) have demonstrated that the transferability of frames across topics is limited. Reimers et al. (2019) have made the case that arguments rarely only evoke one frame and that often multiple aspects are emphasized. In alignment with this observation, Schiller et al. (2021) have operationalized the assignment of frames as a span extraction task rather than as a document classification task. Following up on this, Ruckdeschel and Wiedemann (2022) present a dataset with topic-specific frame classes annotated on token-level.

We directly build on the work of Schiller et al. (2021) and Ruckdeschel and Wiedemann (2022) as a starting point and rely on an argument in which each token is labeled with a corresponding topic-specific frame. This allows us to select the token/spans that have to be modified to switch the frame.

Our proposed approach is inspired by research in XAI, which uses counterfactuals to explain classifier decisions. In the context of XAI, counterfactuals are explanations rooted in counterfactual reasoning. This process entails pinpointing the specific features that, if altered, would result in

[1] https://github.com/phhei/counterfactualREframing

different outcomes or predictions (Miller, 2019). Given this, applying counterfactual approaches to reframing feels intuitive, since changing the frame is conceptually similar to changing a classifier's prediction.

From the literature on counterfactuals, we adopt the idea that suitable metrics can be used to guide the search in the space of potential counterfactuals. Common metrics for selecting an appropriate counterfactual are validity and proximity. A recent paper catalogued up to eight such metrics from contemporary XAI research on generating and evaluating counterfactuals (Verma et al., 2020). In our work, we reuse the metrics related to validity, proximity, and data manifold closeness that are explained in section 3.2.

Counterfactual methods in natural language processing have primarily been used to explain and evaluate sentiment classifiers (Wu et al., 2021; Madaan et al., 2021) or to uncover dataset artifacts (Ross et al., 2021). To our knowledge, their application in the context of reframing is novel, marking a primary contribution of our paper.

3 Methodology

We model the task of reframing as a generative mask-and-replace approach. Given an argument and its frameset $\mathcal{S}$, called source frameset, and a target frameset $\mathcal{T}$, the task is to shift the aspects covered by the argument towards this target frameset $\mathcal{T}$ by rewriting it. Hence, the goal is to remove n_d frames contained in set D to be deleted and add n_a new frames in a set A that are not contained in $\mathcal{S}$. The frameset of the rewritten argument is thus expected to be identical with the target frameset $\mathcal{T} = (\mathcal{S} \setminus D) \cup A$.

Our unsupervised approach is described in Figure 1. In particular, given an argument to be reframed, we apply a sequence tagging model to classify each token into its corresponding frame, relying on the approach proposed by Ruckdeschel and Wiedemann (2022). We then mask each token that has been assigned a frame label that is in the set D. For each masked span, a language model generates an alternative text span which is placed in the corresponding spot, resulting in a new text that is a mixture of original text spans and newly generated text spans.

In order to guide the replacement of a masked span by a span related to set A, we rely on two strategies to increase the ratio of successfully reframed arguments: framed decoding and various output reranking strategies based on the field of counterfactual explanations. We explain these strategies in more detail in what follows.

3.1 Framed Decoding

We follow the proposal of Heinisch et al. (2022) to increase the probability of generating tokens of the target frames to the given argument. In our introductory example, for instance, our goal would be to increase the probability of generating tokens related to an economic frame, such as *affordable* in the example.

For a given frame f, we compute $p(f|v)$, that is the (conditional) probability that if v occurs, it occurs in a text position labeled with frame f. This measures the specificity of v for frame f. At inference time, we modify the logit for each vocabulary element l_v for each target frame $f_t \in \mathcal{T}$ to be added as follows:

$$\tilde{l_v} = l_v + \lambda(\max(l) - \min(l))\ p(f_t|v) \quad (1)$$

where λ is a hyperparameter controlling the degree to which vocabulary elements related to the frames to be targeted are boosted. While a small value of λ yields only a weak boost, a high value strongly boosts tokens related to the frames to be added, potentially leading to output that is unrelated to the input. In order to avoid the repetition of frame-exclusive vocabularies and to aim for frame diversity when multiple frames are applied, we set the repetition penalty to $1 + \lambda$ as proposed by Keskar et al. (2019).

3.2 Reranking strategies

We decode the model using beam search to yield n rewrites of the original sentence. We then apply a re-ranking strategy to select sentences that best align with a quartet of metrics. The first three metrics derive inspiration from the collection presented by Verma et al. (2020) for counterfactuals: i) Frame-Validity, ii) Proximity, iii) Data Manifold Closeness. The fourth metric, iv) Grammaticality and Fluency, is tailored to our specific requirements.

Equation 2 shows how the metrics are aggregated to obtain the final score, which is used to rerank

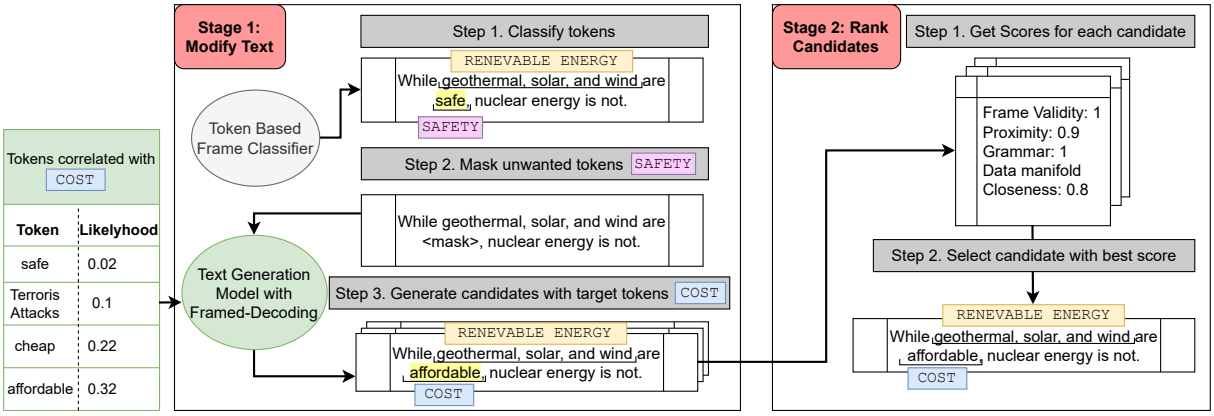

Figure 1: Proposed Reframing Method with Reranking Strategy via Counterfactual Properties

the rewrites in descending order:

$$\begin{aligned}\text{score} = &\ \omega_{\text{validity}} \cdot \text{frame-validity} \\ &+ \omega_{\text{proximity}} \cdot \text{proximity} \\ &+ \omega_{\text{closeness topic}} \cdot \text{data\_manifold\_closeness}_{\text{topic}} \\ &+ \omega_{\text{closeness frame}} \cdot \text{data\_manifold\_closeness}_{\text{frame}} \\ &+ \omega_{\text{grammar}} \cdot \text{grammar}\end{aligned} \tag{2}$$

with ω_m as a weight hyperparameter for metric m.

Frame-Validity The aim of our approach is to generate a rewriting of the given argument that evokes the frames to be targeted. In analogy to the criterion of validity that is used in counterfactual explanation generation to measure the degree to which generated counterfactuals switch a classifiers' prediction, we introduce the analogous *frame-validity* metric that indicates whether the reframing has been successful. For this, we compute a weighted Jaccard similarity between the target frames $\mathcal{T}$ and the frames predicted by the sequence labeling approach $\mathcal{P}$ for the reframed argument, where the weights correspond to the probabilities of the predicted frames:

$$\frac{\sum_{f \in \mathcal{P} \cap \mathcal{T}} p(f)}{\#\mathcal{T} + \sum_{f \in \mathcal{P} \setminus \mathcal{T}} p(f)} \tag{3}$$

Proximity Proximity is used to ensure that the generated counterfactual is semantically close to the original example in counterfactual explanation generation approaches. As we aim for a minimal modification of the argument that effectively reframes the argument, we apply a similar metric in our approach. We aim to maximize the proximity of the generated argument to the original argument, computed by using a Sentence-Bert-model (Reimers and Gurevych, 2019) to embed both sentences and calculate the cosine similarity between them.

Data Manifold Closeness Counterfactual explanation generation approaches aim to generate 'realistic' counterfactuals with a high probability of originating from the actual data distribution. The same holds for reframed arguments, so we transfer the *Data Manifold Closeness* used in counterfactual explanation generation approaches to the reframing task. We aim for reframed arguments to have a strong relation to the desired frames as well as to the issue/topic under discussion. To compute the similarity to the frame and topic, we take the top-k Sentence-Bert embedded neighbors and take the average cosine similarity between those.

Grammaticality and Fluency An important goal is to ensure the grammaticality and fluency of the reframed arguments, so that as a further metric we compute the acceptability of the sentence according to the corpus of linguistic acceptability (CoLA) by Warstadt et al. (2019).

4 Experiment Design

4.1 Dataset

We use the Argument Aspect Corpus (AAC) by Ruckdeschel and Wiedemann (2022) that features manually annotated frame labels for token spans within argumentative sentences. These sentences were drawn from the UKP Sentential Argument Mining Corpus by Reimers et al. (2019), expressing a stance on three major political topics: *m*inimum wage, *n*uclear energy, and *m*arijuana legalization. Since the dataset offers slightly above 1,000 annotated sentences for each topic on the token level, it fits our token-based reframing setting.

4.2 Experimental Settings

As our approach relies on a model that can assign a frame to each token of a given sentence, we reproduce the model proposed by Ruckdeschel and Wiedemann (2022), using the exact same hyperparameters and dataset. We consider the best variant based on `roberta-large` with a sequence tagging head, using the best-performing fine-tuned model on the test data across 5 runs. On the token level, we yield micro-averaged F1 scores across all 12 to 13 frame classes (the topic-specific framesets are defined by Ruckdeschel and Wiedemann (2022)) of 0.63, 0.6, and 0.69 for "nuclear energy", "minimum wage", and "marijuana", respectively.

For the language model generating rewritings for the masked tokens, we rely on the pretrained T5-model variants `t5-small` (60 million parameters) and `t5-large` (770 million parameters) (Raffel et al., 2020) as implemented in the `transformers`-library by Wolf et al. (2020). We mask all token spans with a predicted frame belonging to the frames to be deleted D with placeholders that were used in the masked language pretraining objective of T5. For each placeholder in an incrementing order, T5 generates alternative text spans which we replaced with the placeholders then. Note that T5 does not repeat the input argument while generating. We generate between 4 and 25 tokens per reframed argument candidate, sampling with a temperature of 1.25. To receive n different candidates, we apply beam search with $2n$ beams.

For reranking the candidates, we apply the automatic metrics as proposed in Section 3.2. As the Sentence-BERT model, we rely on `all-MiniLM-L6-v2` (or the more complex model `all-MiniLM-L12-v2` in experiments where `t5-large` was used). For the Data Manifold Closeness, we chose $k = 5$. For the grammar score, we rely on the model `textattack/roberta-base-CoLA` provided by Morris et al. (2020).

Determining the Target Frameset $\mathcal{T}$ Given the frameset $\mathcal{F}$ defined for the debated topic of the argument (Ruckdeschel and Wiedemann, 2022) and the set of all frames $\mathcal{S}$ contained in the argument as predicted by the frame classifier, we randomly delete $n_d \in \{0, 1, 2\}$ frame classes D from $\mathcal{S}$ and randomly add $n_a \in \{0, 1, 2\}$ frame classes from $\mathcal{F} \setminus \mathcal{S}$. In our primarily evaluated reframing setting, we select $n_d = n_a = 1$, exchanging a single frame class in the set of frame classes emphasized by an argument.

Manual study In order to evaluate the reframed arguments beyond using the automatic frame predictions and measurements as proposed in Section 3.2, we conduct a manual study involving three paid annotators, students from the field of (computational) social science.

For the assessment of frames, we use the original well-explored and reviewed guidelines by Ruckdeschel and Wiedemann (2022), including the definition and examples (when given) for each specific frame. In order to ensure a fair evaluation, we hide the original frames of the argument as well as the target frames. Annotators were thus asked to select none or up to five relevant frames evoked by the reframed argument. This is in contrast to studies that ask annotators to confirm whether the reframed argument fits the target frame as a choice between yes, partial, and no (Chen et al., 2021).

For the assessment of grammar and fluency, each annotator had to rate the reframed argument on a Likert scale between 1 (broken/unfinished text) to 5 (perfect fluency and grammar). For the assessment of meaning, each annotator provided two binary labels: one for the preservation of meaning in relation to the original argument and another for the plausibility of the proposed argument as a valuable contribution to the discussion.

On the task of indicating the relevant frames, we obtain a fair inter-annotator-agreement of $\alpha_\kappa = 0.32$ according to Krippendorff's alpha measure, which is comparable to other tasks in the field of argumentation. While we observe an almost perfect agreement in frames that are directly mentioned in the text, e.g. fossil fuels in the first example of Table 3, disagreement occurs in cases of implicit concepts or weakly related implications such as reliable energy when only "special needs by industry" is mentioned. The agreement on the tasks of labeling fluency and grammaticality ($\alpha_\kappa = 0.15$[2]) and meaning ($\alpha_\kappa = 0.16$) are lower due to the subjectivity of these tasks. However, we observe common trends. In terms of grammaticality, we receive constant low grammar scores for obviously broken sentences. Higher deviations are mostly caused by irregular punctuation in which different perspectives are acceptable. In terms of the binary categories related to the meaning, we observe dis-

[2] aligning the annotator-specific mean score across the annotators

agreement on borderline examples and different penalizations of tautologies and repetitions. For example, while all annotators agree regarding the plausibility of the first example in Table 3, they disagree whether the reframed argument is still related to the original argument. Further details on the manual study are provided in Appendix B.

5 Results

We analyze the appropriateness of our reframed arguments along different dimensions. First of all, we evaluate the reframing success of our approach, that is the success of fitting the target frameset $\mathcal{T}$. For this, we compare the frames covered by the reframed argument ($\mathcal{P}$) with the target frames $\mathcal{T}$. Note that $\mathcal{P}$ is predicted by the model of Ruckdeschel and Wiedemann (2022) in the case of the automatic evaluation and annotated by humans in the case of the manual evaluation. We present results with respect to the frames towards three criteria: i) FIT measuring the *target-set-fit ratio*, that is the ratio of instances where $\mathcal{P} = \mathcal{T}$, ii) REM measures the ratio of instances where the unwanted frames are successfully removed, i.e. $\mathcal{P} \cap D = \varnothing$, and iii) ADD measures the ratio of successfully added frames $A \subseteq \mathcal{P}$. These three metrics allow us to judge the reframing validity of our approach. The automatic results covering three topics are presented in Section 5.1. In Section 5.2 we broaden the perspective by also including further metrics that measure other relevant aspects of the reframed argument beyond frame-validity, considering the other automatic metrics introduced in Section 3.2. In addition, we present the results of our manual study in Section 5.3, adding the criteria of meaning preservation and plausibility, evaluating the impact of reranking and framed decoding as well as the impact of the model size of the text-generating model (Section 5.3.1) and the impact of the edit distance between the source frameset $\mathcal{S}$ and the target frameset $\mathcal{T}$ (Section 5.3.2).

We conducted further experiments regarding the impact of the number of rewritings per argument in Appendix A.

5.1 Evaluating Reframing Success

This section evaluates the reframed arguments using `t5-small` by automatically retrieving frames from the rewritten arguments with the task of removing one frame and adding a new frame. In order to exclusively focus on the contained frames while reranking, we set the weights of all counterfactual metrics to 0 in Equation 2 except $\omega_{\text{validity}} = 1$. The automatic results are provided in Table 1. Regarding yielding the target frameset $\mathcal{T}$ as the predicted frameset (FIT), we see an improvement of approximately 4 times by using reranking among 10 rewrites across all three topics. Using nuclear energy as an example topic, the ratio of successful reframing increased from 2.1 to 8.6. Activating framed decoding ($\lambda = 0.5$) improves again the ratios by approximately 6 times (more than an order of magnitude compared to the baseline using a vanilla language model without reranking), yielding ratios of 53.9, 40.1, and 50.9 for nuclear energy, minimum wage, and marijuana. With respect to the ability of the model to remove the frames to be deleted as measured by REM, we observe the same trends but with only comparable minor gains. Vanilla language models are already good at generating replacements that do not share the same frame, having success ratios between 82.6% and 89.7%. Reranking (gaining between 1.9% and 4.4%) as well as framed-decoding (gaining between 6.4% and 11.7%) increases the ratio further, ending with an almost guaranteed frame removal (e.g. 98% for marijuana). Looking at the success rate of adding frames as measured by ADD, we observe major gains using reranking and framed decoding comparable to the FIT analyses, yielding ratios between 45.8% (minimum wage) and 63.4% (nuclear energy).

Note that the results are worse for all topics when decreasing the strength of framed decoding (λ-value) from 0.5 to 0.1, showing the importance of a higher boost of frame-related tokens.

The following example illustrates a common pattern using a high value of $\lambda = 0.5$: Reframing the argument against nuclear energy *"Italy, Belgium, Spain and Switzerland have also principally decided to become nuclear energy-free"* emphasizing the aspect of energy policy ($\mathcal{S}$) towards an argument emphasizing renewable energy ($\mathcal{T}$) results in *"It is essential solar panels wind farms concentrated concentrated in hydro biomass farms to become nuclear energy-free."*, which is barely understandable. The text mentions several technologies for renewable energies to maximize the probability of this particular frame and avoids any names of countries or decision processes to minimize the probability of being labeled with energy policy. This example shows that beyond successfully switching the

	Nuclear energy			Minimum wage			Marijuana		
	REM	ADD	FIT	REM	ADD	FIT	REM	ADD	FIT
MLM (T5$_{\text{small}}$)	84.0	2.9	2.1	82.6	2.5	1.3	89.7	3.0	2.5
⌞**+rerank (10)**	88.4 ‡	10.9 ‡	8.6 ‡	85.1 ‡	9.5 ‡	6.8 ‡	91.6 ‡	12.2 ‡	9.7 ‡
⌞**+frame-dec**$_{\lambda=0.1}$	89.9	20.4 ‡	15.9 ‡	84.1	10.8	8.3	93.0	21.2 ‡	17.8 ‡
⌞**+frame-dec**$_{\lambda=0.5}$	**95.8** ‡	**63.4** ‡	**53.9** ‡	**96.8** ‡	**45.8** ‡	**40.1** ‡	**98.0** ‡	**55.7** ‡	**50.9** ‡

Table 1: Ratios in % of evaluating the reframing success. (‡) significant improvement to the method above with p < 0.005 according to the approximate randomization test with 10.000 resampling steps.

	Nuclear energy		
	∅	FIT	Gram.
MLM (T5$_{\text{small}}$)	56.7	2.1	73.8
⌞**+rerank (10)**	61.1 ‡	8.3 ‡	**85.6** ‡
⌞**+frame-dec**$_{\lambda=0.1}$	62.9 ‡	15.9 ‡	85.3
⌞**+frame-dec**$_{\lambda=0.2}$	**65.7** ‡	38.1 ‡	80.1
⌞**+frame-dec**$_{\lambda=0.5}$	62.7	**53.4** ‡	39.3

Table 2: Scores (0-100) for the different model variants on nuclear energy: Average of all metrics, target-set-fit (FIT), and Grammaticality.

frame, grammaticality and preserving topicality are crucial, so we evaluate our arguments with respect to further criteria in the following section.

5.2 Evaluation including other Reframing Aspects

In order to analyze the appropriateness of reframed arguments beyond the reframing success, we evaluate them with respect to all other metrics introduced in Section 3.2 by introducing an unweighted average of those five (∅), scaling each metric from 0 to 100. However, for reranking, while still regarding frame-validity as the most important metric, we compute an aggregate involving all metrics with weights ω as follows[3]: $\omega_{\text{validity}} = 4$, $\omega_{\text{proximity}} = 1$, $\omega_{\text{closeness topic}} = 1$, $\omega_{\text{closeness frame}} = 0.5$, and $\omega_{\text{grammar}} = 2$ (Equation 2)

The results of the automatic evaluation using again `t5-small` exchanging exactly one frame are provided in Table 2. We observe that the reranking improves every single metric and, hence, the average score. In the case of the topic of nuclear energy, the improvement is 3.4 points, increasing from 56.7 to 61.1. While looking at the different rewrites, we notice that arguments with shorter replacements are preferred on average in order to avoid hallucination and therefore optimize proximity and data manifold closeness while ensuring a high frame-validity. Introducing framed decoding shows a tradeoff between target-set-fit (favoring high λ) and grammaticality/proximity (favoring low λ). The highest target-set-fit ratio (53.4%) is achieved at $\lambda = 0.5$ at the expense of a lower grammaticality (39.3). Conversely, deactivating framed decoding yielded the highest score in terms of grammaticality (85.6) but lowered target-set-fit (8.3%). Thus, framed decoding enforces the target frameset but decreases the (linguistic) coherence, moving the reframed argument away from the original. We find the optimal λ value at 0.2 with a 38.1% ratio of fitting target framesets and a grammaticality score of 80.1. Table 3 shows examples using this setting, containing one successfully reframed argument and two examples of failing to introduce the added frame in the target set.

5.3 Manual Evaluation

To explore the trade-off between target-set fit and linguistic acceptance further, we conducted a manual study with 50 randomly selected arguments derived from the debate on nuclear energy. Once, we exchanged one frame without framed decoding and twice with framed decoding ($\lambda = [0.1, 0.2]$). We automatically selected the best-reframed sentence out of 10 each using the proposed weights in Section 5.2. Table 4 presents the results of the 150 annotated arguments, incorporating the majority vote for frames and meaning and mean values for grammaticality/fluency.

The results of the manual evaluation generally confirm the results of the automatic evaluation. Deactivating framed decoding results in a low target-set-fit (8% of the generated arguments add the new frame, 64% of them remove the deleted frame, and 4% fit the target frameset exactly). However, these arguments have only minor grammaticality/fluency flaws with an average of 3.9, every second preserv-

[3]In an application case such as a dashboard with sliders, a user of the system could select an individual weighting of the different metrics to get different reranked lists.

Original argument	Reframed argument
Just to maintain the current world production of nuclear power, either the oldest, creakiest plants need to be relicensed or a veritable orgy of nuclear construction needs to begin. [RELIABILITY]	*There is a need for fossil oil*, either the oldest, creakiest plants need to be relicensed or a veritable orgy of nuclear construction needs to begin. [FOSSIL FUELS]
The support of nuclear power by government results from special pleading lobbying by the industry. [ENERGY POLICY]	The support of nuclear power by the *industry* results from special *needs by the industry*. [RELIABILITY]
Prospects for nuclear energy as an option are limited, the report found, by four unresolved problems: (1) high relative costs; (2) perceived adverse safety, environmental, and health effects; (3) potential security risks stemming from proliferation; and (4) unresolved challenges in long-term management of nuclear wastes. [COSTS], [ACCIDENTS/SECURITY], [ENVIRONMENTAL IMPACT], [HEALTH EFFECTS], [WASTE]	Prospects for nuclear energy as an option are limited, the report found, by four unresolved problems: (1) high relative costs; (2) perceived adverse safety, *safety*, and health effects; (3) potential security risks stemming from proliferation; and (4) unresolved challenges in long-term management of nuclear wastes. [COSTS], [ACCIDENTS/SECURITY], [HEALTH EFFECTS], [WASTE], [TECHNOLOGICAL INNOVATION]

Table 3: Examples using t5-small+rerank (10) with framed decoding ($\lambda = 0.2$), removing and adding one frame class

	w/o λ	**$\lambda = 0.1$**	**$\lambda = 0.2$**
Success of reframing (%)			
REM	64	82	**82**
ADD	8	8	**26**
FIT	4	4	**18**
Grammar/ Fluency (1-5)			
∅	**3.9**	**3.9**	3.7
Meaning (%)			
Preservation	**50**	48	40
Plausibility	60	**68**	60

Table 4: Results of manual evaluation (t5-small + rerank (10)), debating *nuclear energy*

	t5-small (10)		**t5-large (10)**	
	w/o λ	$\lambda = 0.2$	w/o λ	$\lambda = 0.2$
Frame-FIT (%)	4	**18**	6	**18**
Grammar (1-5)	3.9	3.7	4.1	**4.1**
Preservation (%)	50	40	**62**	60
Plausibility (%)	60	60	70	**72**

Table 5: Results of the manual study considering two model variants, debating *nuclear energy*

ing the meaning of the original argument, and 60% of which are plausible. Activating the framed decoding again shows a similar λ influence with a sweat-spot of $\lambda = 0.2$, yielding a high target-set-fit (18%) and generating well-formulated arguments (3.7) while preserving meaning (40%) and plausibility (60%).

In comparison to the automatic evaluation results shown in Table 2, we notice a significant drop by $\approx 50\%$ in the target-set-fit ratio. This discrepancy can be primarily attributed to the use of the same classifier for both the automatic evaluation and the classification of tokens with frames. This classifier plays a crucial role in identifying the text segments that need to be replaced to achieve a new target frameset. As a consequence of this setup, incorrect frame predictions that occur outside the replaced text segments go unnoticed in the automatic evaluation but are detected in the manual evaluation.

5.3.1 Impact of model size

To analyze the impact of using a larger model (namely `t5-large`), we expanded our manual annotation study by 50 reframed arguments for each hyperparameter setting. Table 5 shows the results.

Regarding the target-set-fit ratio, we observe similar performances, yielding only 4% and 6% for `t5-small` and `t5-large`, respectively, without framed decoding. While `t5-small` is better in avoiding the removed frame class (64%) but

not successful and targeting the added frame class (8%), `t5-large` is worse in removing (58%) but better in adding (16%). Due to the higher model complexity, `t5-large` is better at generating context-fitting replacements, having a higher chance to restore the masked text part but also to uncover new aspects, while `t5-small` tends to generate more general and debate-unspecific replacements, resulting in less meaning preservation (dropping from 62 to 50%) and plausibility (from 70% to 60%).

Nevertheless, activating the framed decoding process with $\lambda = 0.2$ reduces the impact of model size regarding the framing. Both text-generating models produce a target-set-fit ratio of 18%, demonstrating the success of our decoding strategy being insensitive to model size. However, `t5-large` shows a better performance on selecting linguistically fitting tokens which leads to comparable ratings in grammaticality ($\varnothing 4.1$), meaning preservation (60%) and plausibility (72%). Here, `t5-small` starts to generate clearly ungrammatical or unfitting text replacements in some cases.

5.3.2 Evaluating reframing on multiple frames

Up to this point, our focus has been on the task of reframing involving the replacement of a single frame class within an argument in a multilabel setting. Next, we experimented with removing and adding none or multiple frame classes simultaneously, exclusively relying on arguments covering at least two frame classes. Due to the increasing complexity, we used `t5-large` with activated framed decoding ($\lambda = 0.2$), again reranking among 10 candidates per argument. The manual analysis incorporated 50 reframed arguments, once for removing 1 frame class (deframing) and once for exchanging 2 frame classes (extended reframing).

Increasing the edit distance between the source frameset $\mathcal{S}$ and target frameset $\mathcal{T}$ increases the task difficulty. With deframing, we achieve a target-set-fit of 24% (yielding 66% reframed arguments without the removed target frame). By exchanging 1 target frame class we measure a target-set-fit of 18% while finally dropping to 8% by exchanging 2 target frame classes due to the major changes needed to achieve the complex changes between $\mathcal{S}$ and $\mathcal{T}$. The challenge of this extended reframing is also reflected by the other three manual metrics but still yielding an average grammar score of 3.9, a ratio of 42% in meaning preservation, and a ratio of 64% in terms of plausibility.

6 Conclusion

We have proposed an unsupervised approach to argument reframing, which takes inspiration from approaches to counterfactual explanation generation in the sense that we transfer and adapt metrics used in counterfactual generation to implement a reranking strategy for reframed arguments. We use an LLM to replace text spans that were tagged by a token classifier with a frame to be deleted by tokens that are associated with the frame to be added.

Our automatic and manual evaluation demonstrates that the combination of framed decoding and reranking, utilizing metrics such as frame-validity, proximity, data manifold closeness, and grammaticality, outperforms a vanilla LLM baseline by nearly an order of magnitude in terms of reframing success. Furthermore, by showing a tradeoff between tailoring the rewritten argument to the target frameset and yielding a plausible and grammatically correct argument, we identified a sweet spot in the strength of framed decoding yielding across two different language generation model sizes.

Acknowledgements

This work has been funded by DFG within the project ACCEPT, which is part of the priority program "Robust Argumentation Machines" (RATIO), and the project B01 within the TRR 318 "Constructing Explainability".

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. 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 2922–2932, Hong Kong, China. Association for Computational Linguistics.

Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

Processing (Volume 2: Short Papers), pages 438–444, Beijing, China. 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.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2021. Controlled neural sentence-level reframing of news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2683–2693, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Claes de Vreese. 2005. News framing: Theory and typology. *Information Design Journal*, 13:51–62.

Philipp Heinisch, Anette Frank, Juri Opitz, and Philipp Cimiano. 2022. Strategies for framing argumentative conclusion generation. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 246–259, Waterville, Maine, USA and virtual meeting. 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.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13516–13524. AAAI Press.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Xinyi Mou, Zhongyu Wei, Changjian Jiang, and Jiajie Peng. 2022. A two stage adaptation framework for frame detection via prompt learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2968–2978, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

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, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Mattes Ruckdeschel and Gregor Wiedemann. 2022. Boundary detection and categorization of argument aspects via supervised learning. In *Proceedings of the 9th Workshop on Argument Mining*, pages 126–136, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

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.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. 2020. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening

the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. 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 6707–6723, Online. Association for Computational Linguistics.

A Analysing the impact of the number of generated candidates for reframing

This section presents a concise analysis of how the quantity of rewrites impacts the quality of the best-reframed argument considering the automatic reranking with weights proposed in Section 5.2, measured by the target-set-fit ratio and the average score of the reranking metrics.

Investigating the debate of "nuclear energy", Figure 2 illustrates a pattern wherein an increasing number of rewrites monotonously increases both metrics across all models since additional rewrites potentially outperform the choice among fewer rewrites, but can not worsen the metrics based on the best argument after reranking. However, the curves flatten with an increasing number of rewrites, representing a stochastic principle of sampling from an ordered distribution. Since we apply sampling at decoding time itself, every language model has the capability to generate every text that maximizes the automatic metrics (100%). Hence, our distribution contains this optimal text which has to be sampled assuming access to infinite rewrites. However, this is not practicable, raising the question of the probability mass of the "good" rewrites. Here, we observe that framed decoding shifts the probability mass significantly, leading to better solutions at fewer rewrites compared to instances where framed decoding is not employed.

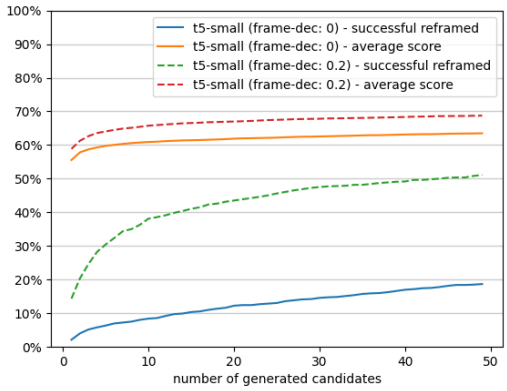

Figure 2: Influence of the number of rewrites debating "Nuclear energy" (`t5-small`)

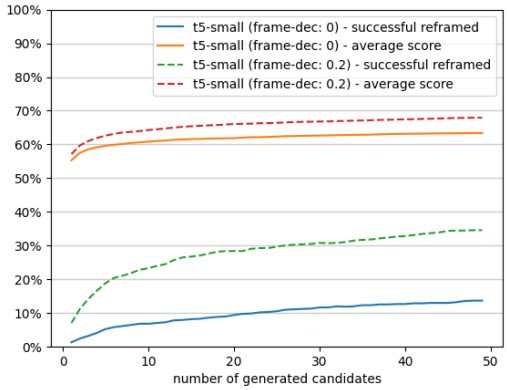

Figure 3: Influence of the number of rewrites debating "Minimum wage" (`t5-small`)

The observation holds for other topics as well. Figure 3 for the topic "minimum wage" shows a similar relation between the metrics and number rewrites, having only a smaller slope of increment. Looking at Table 1, we see that this topic yields the smallest ratio of successfully reframed arguments on average, suggesting more complex frame classes. Nevertheless, the same stochastic principles apply here, observing the same trends.

B User Interface of the Manual study

Figure 4 shows the user interface of our manual study. Annotators were shown one argument at a time and were asked to rate the mentioned frames, the fluency, and the meaning. Each frame class is

Frame	α_κ
Accidents/security	0.361
Costs	0.462
Energy policy	0.133
Environmental impact	0.362
Fossil fuels	0.286
Health effects	0.498
Public debate	0.107
Reliability/efficiency	0.173
Renewables	0.386
Technological innovation	0.209
Waste	0.454
Weapons	0.457
Overall	0.324

Table 6: Inter-annotator agreements (Krippendorff's Alpha) of the manual user study, topic *Nuclear energy*

described adapting the original descriptions[4], once by hovering the frames and once in the guidelines at the bottom of the page, containing examples as well. The annotators answered the questions sample by sample independently from each other.

C Replacing T5 with larger prompt-based Large Language Models

Recent advancements in prompt-based Large Language Models, such as chatGPT or the successor GPT-4 (OpenAI, 2023), show wide applicability for many NLP tasks in a few- or even zero-shot setting. To test the potential for usage as reframing models, we used a prompt[5] to test the performance of GPT-4 on some selected examples shown in Table 3:

1. Just to maintain the current world consumption of fossil fuels, either the oldest, most depleted fields need to be rejuvenated or a significant surge in new drilling needs to begin."

2. "The support of nuclear power by the government results from its reliability in the industry.

3. "Prospects for nuclear energy as an option are limited, the report found, by four unresolved problems: high relative costs; perceived adverse safety and technological challenges; health effects; potential security risks stemming from proliferation; and unresolved challenges in long-term management of nuclear wastes."

While the first look at the reformed arguments is promising (introducing related phrases towards the frame class which should be added in all arguments), we see critical drawbacks using GPT-4. Although the parts marked as "fit the target frame set" of the reframed arguments align with the original argument, GPT-4 failed to keep them completely unchanged and, hence, perform more changes than necessary, leading to less controllability. Furthermore, with respect to the automatic frame class prediction, GPT-4 often fails to reframe successfully. In the presented examples, GPT-4 successfully added only once the new frame class and failed two times to remove the frame class that should have been discarded. GPT-4 shows also a dependency on descriptions of the frame classes, e.g. to guide the second example towards "reliable energy" rather than "reliability" in general. All in all, GPT-4 alone without further guidance as provided by framed decoding or reranking is not suited to support the type of minimalistic reframing that we are targeting. However, using these two techniques to introduce framing capabilities in an unsupervised manner, we require only a general language understanding of the underlying generative language model. Using a much larger prompt-based model with more capabilities is not necessarily beneficial here. In order to keep the requirements for the computational resources realistic, especially with respect to a beam search using up to 100 beams in order to yield a comprehensive search space for counterfactual reranking, we consider T5 as the model of choice for this paper.

[4] https://zenodo.org/record/7525183/files/AAC_NE_Guidelines.md?download=1

[5] You are an assistant for reframing sentences that are tagged with specific aspects. The current topic is nuclear energy and the tags show which aspects of the topic the tokens belong to. You will be given a sentence with a set of initial aspects and your task is to perform minimal changes on the sentence to reframe it into a new target set of aspects, without changing words that are not labeled to an aspect that needs to be removed. The new aspects are general topics and not the words that need to be included.

This is a tagged sentence with the aspects [SOURCE SET] and the target set is [TARGET SET]

[Sentence with annotated labels]

Now perform minimal changes to this sentence to achieve a reframed sentence that has the target set as annotated aspects. Try to keep the sentence as close to the original one and change only what is necessary. The fewer changes the better. Keep the tokens that are not related to the reframing the same, i.e. don't remove unnecessary tokens if they are not related to an aspect that needs to be removed. Write the new sentence without aspect classifications but just as a plain sentence.

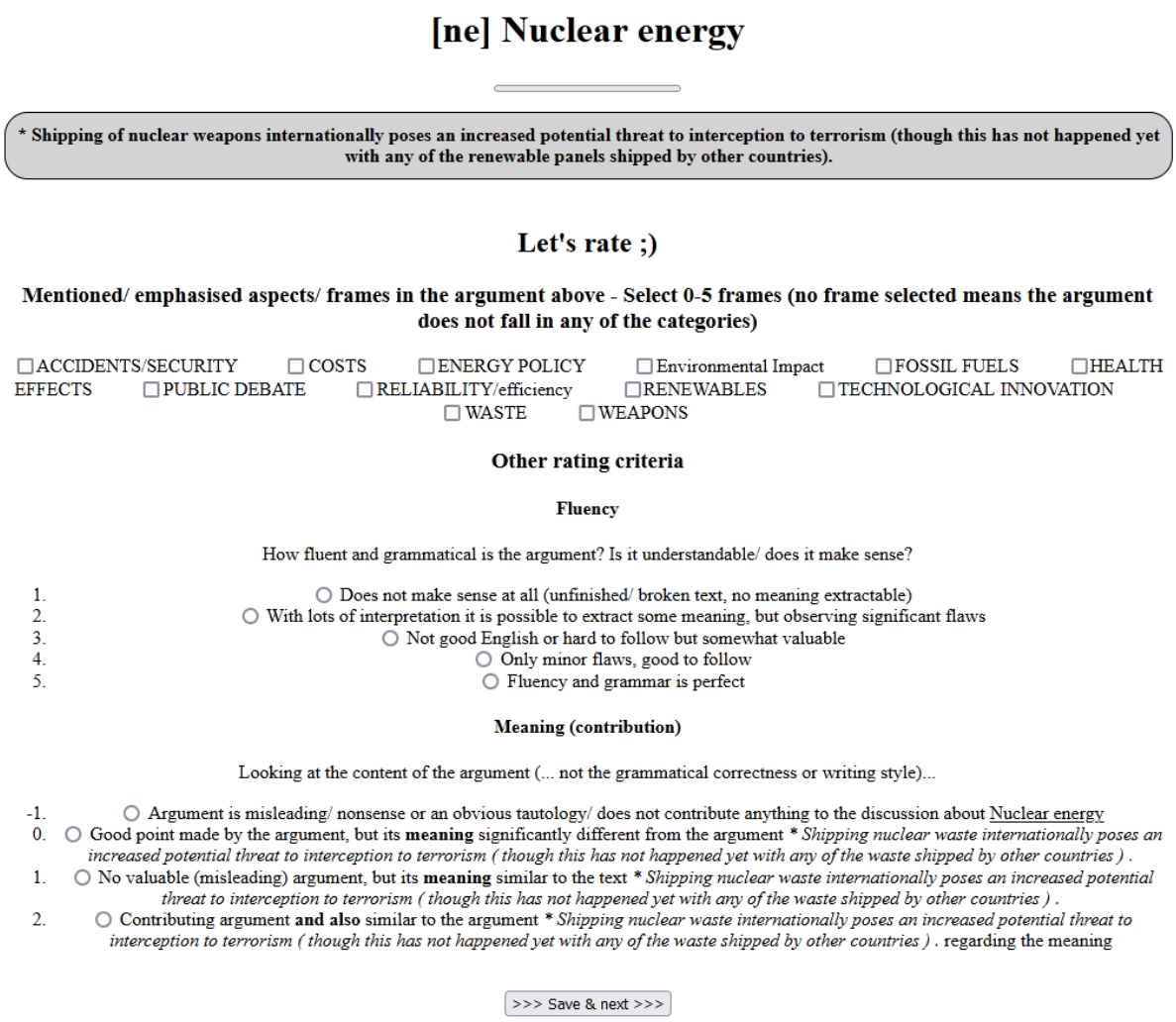

Figure 4: Screenshot of part of the annotator interface of the manual study