Strategies for Framing Argumentative Conclusion Generation

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

Anette Frank
Heidelberg University
frank@cl.uni-heidelberg.de

Juri Opitz
Heidelberg University
opitz@cl.uni-heidelberg.de

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

Abstract

Generating an argumentative conclusion from a set of textual premises is a challenging task, due to a large range of possible conclusions. In order to provide a conclusion generation model with guidance towards generating conclusions from a certain perspective, we explore the impact of conditioning the model on information about the desired framing. We experiment with conditioning generation via generic frame classes as well as with so-called issue-specific frames. Beyond conditioning the model on a desired frame, we investigate the impact of strategies to further improve the generated conclusion by i) an informative label smoothing method that dynamically smooths one-hot-encoded reference conclusion vectors as a regularization mechanism, and ii) a conclusion reranking strategy based on reference-less scores at inference time. We evaluate the benefits of our methods using metrics for automatic evaluation complemented with an extensive manual study. Our results show that frame-guided conclusion generation is beneficial: it increases the ratio of valid and novel conclusions by 23%-points compared to a baseline without frame information. Our work indicates that i) by injecting frame information, conclusion generation can be directed towards desired aspects and ii) at the same time it can be manually confirmed to yield more valid and novel conclusions.

1 Introduction

Argument mining enables systems to automatically retrieve (Wachsmuth et al., 2017a), analyse (Becker et al., 2020), classify (Trautmann et al., 2020), rank (Wachsmuth et al., 2017b) or summarize (Bar-Haim et al., 2020) arguments on a controversial topic. In line with growing amounts of user-generated argumentative content, this field is intensely researched and bears the potential to support humans in deliberation (Fromm et al., 2019).

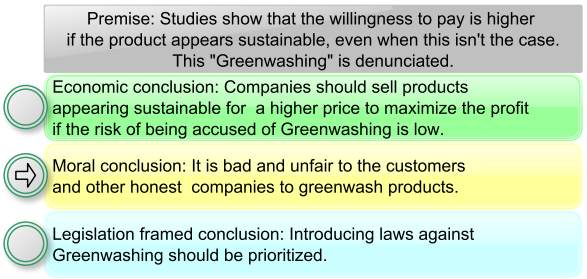

Figure 1: Example argument. All three conclusions are appropriate, but are framed in different ways.

An argument can be conceptualized as a pair of premise(s) and a conclusion. At the core of an argument lies the inferential link between the premises (evidences) and the conclusion. Current systems capture this inferential link only to a limited extent, e.g. by predicting whether a premise supports or attacks a given claim (Cocarascu et al., 2020). While approaches such as Paul et al. (2020) establish chains of background knowledge that characterize the link between premises and conclusions, such methods are limited to analyzing *existing* arguments. To better understand whether computational systems are able to draw inferences from premises towards conclusions, in this work we study the problem of automatic conclusion generation. Being able to automatically *generate* conclusions bears great potential: not only could we retrieve arguments and make their unstated conclusions explicit – such a method would also allow us to generate novel arguments in a debate, thereby supporting deliberation – by raising novel conclusions from different perspectives.

Yet, the process that infers a conclusion from a set of premises is underspecified, since different conclusions may be drawn from a set of premises, depending on different viewpoints. An example is illustrated in Figure 1. It shows the importance of being able to reflect on a topic under discussion from various perspectives (de Vreese, 2005).

Proceedings of the 15th International Conference on Natural Language Generation, pages 246 - 259
July 18-22, 2022

Two approaches have been used to describe the different perspectives or *"framings"* of a discussion. Several authors (Neuman et al., 1992; Boydstun et al., 2014) have proposed to work with a fixed set of manually defined frames, so-called *generic frames*. Others, adopting a more open approach (de Vreese, 2005), have proposed to rely on a vocabulary of *issue-specific frames* that vary from debate to debate, are more fine-grained, and can be provided by users to cluster their arguments in a certain debate (Ajjour et al., 2019). Building on these two notions of *framing*, we investigate which type of frame information is most effective to guide a conclusion generation model.

Previous work has attempted to reconstruct a missing conclusion by identifying the "main target" in the premises (Alshomary et al., 2020). Other work has made use of pretrained sequence-to-sequence transformer language models fine-tuned on argumentative datasets (Syed et al., 2021; Opitz et al., 2021; Gurcke et al., 2021). However, the question of how to tailor a generated conclusion to a particular frame has not been systematically explored, a gap that we address with this paper. Our contributions are the following[1]:

i) We present a framework and method based on autoregressive transformer-based decoding to study how the generation of (textual) conclusions can be controlled by integrating information about the desired frame as input. We explore different frame granularities separately and in combination: generic frames as defined by Boydstun et al. (2014) and issue-specific frame labels.

ii) We present results on the issue-specific frames dataset by Ajjour et al. (2019), showing improvements resulting from conditioning on a desired frame, through i) automatic evaluation, as well as ii) a study relying on human annotators rating *validity, novelty* and *frame relatedness* of the conclusion.

iii) We investigate additional strategies to guide the conclusion generation model towards selecting an appropriate conclusion, using a label-smoothing method applied at *training time*, and two strategies (frame-sensitive decoding and conclusion reranking) applied at *inference time*. These additional methods yield further improvements, while highlighting an interesting trade-off between validity and novelty of the generated conclusions.

[1]Our code is available on GitHub: phhei/ConclusionGenerationWithFrame

2 Related work

While massive amounts of user-generated arguments are available in various debate portals or writing platforms, these arguments are often incomplete, missing an explicitly stated conclusion or lacking essential premises. Such omissions are frequent and often due to rhetorical reasons (Rajendran et al., 2016; Becker et al., 2021). However, arguments lacking an explicit conclusion create challenges for downstream processing tasks (Opitz et al., 2021; Alshomary et al., 2020; Gurcke et al., 2021). Thus, prior work has investigated approaches to make conclusions explicit.

First approaches in this direction attempt to extract missing parts by copying from similar or related arguments, or by applying common, handcrafted argument patterns (Rajendran et al., 2016; Reisert et al., 2018). Yet, these approaches are limited due to the variety of human argumentation and do not generalize well to novel topics.

More recent works leverage sequence-to-sequence transformer language models: Syed et al. (2021) is the first approach known to us that relied on transformer models to generate conclusions given premises. They relied on a pretrained BART model showing that it is able to create premise-related text. However, their manual study shows that 14-36% of the generated conclusions are valid, e.g. by rephrasing the premise, but only 4-6% are informative. Opitz et al. (2021) also show that state-of-the-art fine-tuned transformer language models processing plain premises tend to generate conclusions lacking in novelty or validity, and proposed ways to assess their novelty and validity using AMR-based similarity metrics. Finally, Gurcke et al. (2021) explored whether the sufficiency of conclusions can be assessed with BART, and find problems with insufficient reference conclusions – with ensuing challenges in generating and evaluating valid and novel conclusions.

Prior work also investigated whether the quality of generated conclusions can be improved by conditioning a language model exclusively on topic and frame. Schiller et al. (2021) show that claims generated by such a conditional transformer language model are in general of high quality.

However, none of the approaches mentioned so far has attempted to directly control the framing of

a conclusion by conditioning the model via a given premise *and* the desired frame, a gap we close in this paper. We investigate different ways of encoding the frame and experimentally investigate the impact of these guides using automatic and human evaluations.

3 Datasets

To study the impact of controlling conclusion generation by conditioning on the desired frame, we rely on two datasets. One is the Media-Frames dataset, which relies on an inventory of 15 generic frames originally proposed by Boydstun et al. (2014). The second dataset, produced by Ajjour et al. (2019), does not rely on a fixed set of frames, but on user-provided frames – so-called issue-specific frames. Details of both datasets are given below.

3.1 Media-Frames dataset

The Media-Frames dataset by Card et al. (2015) consists of 17,826 newspaper articles on three policy issues (*immigration*, *smoking* and *same-sex marriage*) annotated with the generic *Media Frames* defined by Boydstun et al. (2014). The set of *Media Frames* contains 15 different frame classes: i) Economic, ii) Capacity and resources, iii) Morality, iv) Fairness and equality, v) Legality, constitutionality and jurisprudence, vi) Policy prescription and evaluation, vii) Crime and punishment, viii) Security and defense, ix) Health and safety, x) Quality of life, xi) Cultural identity, xii) Public opinion, xiii) Political, xiv) External regulation and reputation, as well as xv) Other. The annotation of frame information was performed in several rounds by selecting text spans and assigning them to one of the 15 frame classes, which yielded an inter-annotator-agreement between 0.29 and 0.6 according to Krippendorff's α. To increase the reliability of the data, we rely on only those instances for which at least two annotators agree on the corresponding frame. The resulting subset contains 21,206 samples.

3.2 Argument dataset with issue-specific frames

Ajjour et al. (2019)'s dataset contains 12,326 arguments that were annotated with user-generated issue-specific frame labels – tags that can serve to cluster arguments in a debate to better overview the controversial aspects. The data is crawled from Debatepedia[2] and consists of 365 different topics. In total, the label set comprises 1,623 different frames labels. Out of these, only 330 occur in two or more topics, which indicates that there is a substantial long tail of labels that occur only a few times.

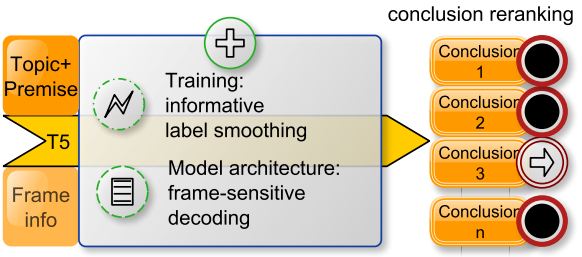

Figure 2: Overview of contributions: Frame-sensitive conclusion generation by frame-sensitive decoding, informative label smoothing and conclusion reranking.

4 Methods

We rely on a sequence-to-sequence encoder/decoder architecture that encodes the topic and the premise, and autoregressively decodes the conclusion. We examine whether and how the generation can be conditioned by enriching the input with information about the desired frame. We investigate i) the explicit encoding of frames as part of the input (4.1) and ii) injection of prior generic frame knowledge by adjusting the output of the language model (4.2). Moreover, we also propose more fine-grained methods: iii) an informative label smoothing training technique and iv) a conclusion reranking approach (4.3). The label smoothing approach attempts to push the model to generating a conclusion that is specific for the given desired frame, while the conclusion reranking method reranks potential conclusion candidates using shallow and argumentation-inspired metrics.

4.1 Explicit encoding of frames

To condition conclusion generation on a frame, we encode the frames (issue-specific and generic frames) as part of the input, as pictured in Figure 3. The input to the transformer model uses additional separators and looks as follows:

```
summarize
  [T] topic [/T]
  [Fis] issue-specific frame [/Fis]
  [Fgm] generic frame (argument) [/Fgm]
  [Fgi] generic frame (conclusion) [/Fgi]:    premise.
```

[2] http://www.debatepedia.org

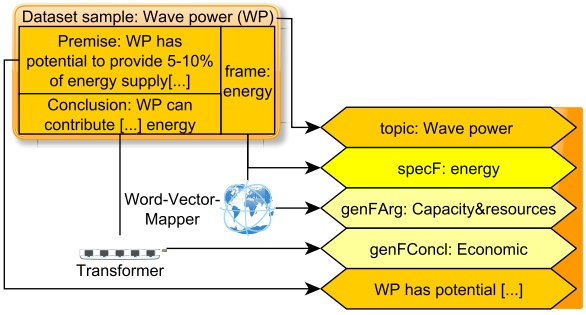

Figure 3: Different input parts for the conclusion-generating language model showing an abbreviated sample of the dataset of Ajjour et al. (2019)

Hereby, *topic* is the debate title as contained in the dataset of Ajjour et al. (2019). The *issue-specific frame* (specF) is the frame label as described by users for each argument in this dataset. For determining the *generic frame (argument)* (genFArg), we map each generic frame class label and the issue-specific frame into a low-dimensional semantic vector space by semantically aggregating the word-vectors as proposed by Heinisch and Cimiano (2021). We select the generic frame label that is closest in this vector space to the given issue-specific frame. Finally, we propose a second approach to inferring a generic frame denoted by *generic frame (conclusion)* (genFConcl) by using a transformer model trained on the Media-Frames dataset (Appendix A.1) that predicts the corresponding frame for the conclusion.

4.2 Frame-sensitive decoding

Our goal is to increase the likelihood that a generated conclusion is frame-specific. For this, we increase the probability that, at decoding time, the model outputs tokens that are associated with the given frame. To achieve this, we follow a finding of Naderi and Hirst (2017) who measured a correlation between particular uni- and bigrams and certain generic frames in the Media-Frames dataset. For example the $-sign often occurs in an economically framed text. We can use this frequencies to inject frame-specific prior knowledge by adjusting the output logits o of the transformer. With this modification we can directly influence the sequence-to-sequence decoding, as shown in equation (1),

$$o'_v = \frac{o_v}{2} + o_v \left(h(o)_v \frac{\log(tf_{D_f}(v)+1)}{\max_{\tilde{v}\in V} \log(tf_{D_f}(\tilde{v})+1)} \right)$$
$$\forall v \in V, o \in \mathbb{R}^{|V|}, h \mapsto [0,1]^{|V|} \quad (1)$$

where v is a vocabulary element, $h(o)$ a parametrizable function that maps the logit values to a range of $[0,1]$, and $tf_{D_f}(v)$ the term frequency of v in documents framed with the generic frame f. Specifically, we set the new output logit o'_v for v to half of the model's logit output, and add this logit's value scaled by its normalized frame-frequency in combination with the overall predicted logits. In this way, a higher frequency of a given word v in frame f results in a higher added value. As a result, we expect the model to prefer generating tokens that are likely to occur in the desired frame f, while dispreferring tokens that are unlikely to occur in texts framed with f.

4.3 Additional strategies to boost frame-sensitive conclusion generation

As a further option to conditioning the autoregressive generator to an explicitly encoded frame in the input and including frame-relevant word knowledge, we now analyse the impact of two strategies that aim to move the generated conclusion closer to the reference conclusions. Specifically, we apply an *informative label smoothing technique* and a *conclusion reranking strategy*.

Informative label smoothing During fine-tuning the language model for conclusion generation, we apply a regularization technique proposed by Szegedy et al. (2016) that modifies the computation of the cross-entropy loss by smoothing each one-hot-encoded conclusion token vector $\vec{y}$, transforming it into an token vector $\vec{y'}$ that distributes part of the probability mass to the whole vocabulary. Given a smoothing strength parameter $\lambda \in [0,1]$ and the token sequence of the reference conclusion $c = \{w_1, \cdots, w_n\}$, the one-hot-encoded vector $y_{\vec{w_i}}$ for each token w_i is transformed as follows:

$$y'_{\vec{w_i}} = \left(\frac{\lambda}{V}, \cdots, 1-\lambda+\frac{\lambda}{V}, \cdots, \frac{\lambda}{V} \right) \quad (2)$$

While λ is fixed for all tokens w_i in the approach of Szegedy et al. (2016), we propose an ex-

tension that uses on a token-specific λ' that multiplies λ by two token-specific factors. The first factor (controllable by $\delta \in [0,1]$) scales λ proportionally to the term frequency $tf(w_i)$. Thus, the more frequent the token w_i is, the higher λ' and thus the more is the output reference vector spread and further away from a one-hot encoded vector. The second factor (controllable by $\psi \in [0,1]$) scales λ inverse proportionally to the frame-specific term frequency $tf_{D_f}(w)$ with which the token occurs in frame f. Thus, the more frequent the token w_i in the frame f, the lower the value of λ' and thus the more is the distribution centered on token w_i[3]. Overall, we adjust the smoothing strength for each w_i as follows:

$$\begin{aligned}&\lambda'(w_i) = \lambda \\ &* \left(1 - \delta + \left(2\delta \frac{\log(tf(w_i)+1)}{\max_{v \in V} \log(tf(v)+1)}\right)\right) \\ &* \left(1 - \psi + \left(2\psi \left(1 - \frac{tf_{D_F}(w_i)}{\max_{v \in V} tf_{D_F}(v)}\right)\right)\right)\end{aligned} \quad (3)$$

Conclusion reranking: selecting the most appropriate conclusion We explore a conclusion reranking strategy inspired by Hua and Wang (2020), to choose a proper conclusion among different beam search traces. Given a set of conclusion candidates $\mathcal{C} = \{c_1, c_2, \cdots, c_n\}$ and a set of automatically calculated reference-less scores $\mathcal{S}(c) = \{s_1(c), \cdots, s_m(c)\}$ for each candidate c, we select the conclusion c' that maximizes a weighted linear combination of the reference-less scores as indicated in the following equation:

$$\begin{aligned}&c' = \arg\max_{c \in \mathcal{C}} \textstyle\sum_{i=1}^{m} \omega_i s_i(c) \\ &\text{with } \omega = (\omega_1, \cdots, \omega_m) \\ &\text{as a fixed pretrained scalar vector.}\end{aligned} \quad (4)$$

We formalize the problem of optimizing the ω-vector as a linear optimization problem [4] in an additional Ω-optimization split, determining the ω-vector by the following equation that minimizes the gap between the reference-less metric scores $\mathcal{S}$ and r metric scores $\tilde{\mathcal{S}}$ using the reference:

$$q_{min}(\Omega) = \sum_{c \in \Omega} |\sum_{i=1}^{m} \omega_i s_i(c_i) - \sum_{h=1}^{r} \frac{\tilde{s_h}(c_h)}{r}| \quad (5)$$

[3]To avoid too small text corpora of a particular frame, the second factor requires generic frame information in the input, otherwise the second factor is disabled

[4]We also tried more complex learning models, for example SVMs. However, more complex models did not outperform the linear regression on average, while increasingly lacking in interpretability.

Examples of metrics which do not consider the reference are listed in A.2.

5 Experiments and Evaluation

In our experiments we investigate the impact of conditioning the generation of an argumentative conclusion, given a topic and premise, on different frame labels provided as part of the input, in addition to our semantic- and frame-sensitive model adjustments. We explore the influence of different types of frame information – the original issue-specific frame labels and the derived generic frame classes – combined with our model variants. We perform automatic evaluation of the generated conclusions, relying on similarity to the reference conclusion as evaluation measure. We also carry out a comprehensive manual evaluation in which annotators rated the validity and novelty of generated conclusions, as well as their closeness to the desired frame.

5.1 Experimental Setup

We follow previous work by Opitz et al. (2021) and rely on T5 (Raffel et al., 2020) (large version) as transformer language model, as implemented in the huggingface library (Wolf et al., 2020).

We use the argument dataset by Ajjour et al. (2019) (Section 3.2), which we divide into splits of 80%, 10%, 5% and 5% of the samples for training, development, Ω-optimization split for conclusion reranking and test, respectively, without overlapping topics.

We test different frame configurations, including subsets of our three frame specifications. The abbreviation *specF+genFArg*, e.g., symbolizes a fine-tuned model that receives the issue-specific and generic frame (argument) as frame information for each sample.

Training: We train between 2 and 12 epochs, where we stop the training process after the validation loss increases in two consecutive epochs. We use an initial learning rate of 2e-4 and decrease it during the training in a step-wise fashion with a factor of 0.975 every 32 steps including a minor weight decay of 1e-7. For our frame-sensitive decoding strategy that uses the MediaFrames dataset as prior knowledge base, we set the function h in Equation 1 to the softmax function. For our informative label smoothing we experimented with different parameters and got strong results with a general label smoothing factor of

$\lambda = 0.1$, a dynamic smoothing factor of $\delta = 0.4$, and a generic frame smoothing factor of $\psi = 0.5$ (Equation 3). The frame-sensitive decoding and the frame-sensitive component of the informative label-smoothing considers the generic frame (preferred from argument) part in the input. If no generic frame information is given in the input, these two frame-sensitive adjustments are deactivated.

Inference: We apply nucleus sampling as proposed by Holtzman et al. (2020) (considering tokens covering 92.5% probability, but max. 50 different tokens, by five beams or twelve in case of conclusion reranking). The temperature is set to 0.75 or 1.1 in the case of conclusion reranking to increase the word diversity. For conclusion reranking, we developed and considered a variety of reference-less scores. We consider shallow surface cues of the conclusion candidate, such as the length (also in ratio to the premise length), the ratio of stop words, the existence of conclusive trigger words such as "should" for normative conclusions, and also non-shallow metrics. To measure the grammaticality of the generated conclusion, we use the GRUEN-score (Zhu and Bhat, 2020). Furthermore, we check deep argumentative characteristics, for example the argumentative relation between the premise and the conclusion, the BERTscore between premise and conclusion candidate, to avoid copies or completely unrelated conclusions, as well as whether the generated conclusion candidate matches the desired frame, if available. We list and further describe all used reference-less metrics in the Appendix A.2. We test two different variants of conclusion reranking. The first optimizes the aggregation of the ROUGE-1, ROUGE-L, as well BERTscore, and thus the similarity to the reference conclusion on the Ω-optimization split, called *frame-insensitive conclusion reranking*. The second *frame-sensitive* variant in addition optimizes the automatic frame-relatedness scores of the selected conclusions on the Ω-optimization split.

Evaluation: We rely on a mixture of automated scores, such as the token-based ROUGE-score and the BERTscore[5] (Zhang et al., 2020) measuring the semantic similarity between generated and reference conclusions.

[5]rescaled f_1, using the 18th layer of `microsoft/deberta-large-mnli` without an idf-weighting

Evaluating the generated conclusions with respect to their references using automatic metrics might, however, penalize valid conclusions that differ substantially from the reference. We therefore also perform manual evaluation with human annotators on 30 randomly selected arguments from the test-split[6]. The annotators were paid for their work and are not authors of the paper. Each reference conclusion, random and generated conclusion is annotated three times by the same three annotators in three consecutive rounds with respect to the following dimensions: (1) **Validity**: Is the conclusion justified based on the premise?, (2) **Novelty**: Does the conclusion contain premise-related novel content that is not part of the premise?, and (3-4) **issue-specific frame / generic frame (argument)**: Is the conclusion directed towards the given frame? To avoid different scale interpretations, we allow only the answers {yes, no, can't decide}. In an additional pairwise setting, presenting two conclusions, we ask whether one (and if so, which) conclusion is better in view of the rated aspect. We hide the source of the presented conclusions (reference, random or the generating model configuration) to avoid bias. Further details on the manual study are given in Appendix A.3.

5.2 Results

Impact of conditioning conclusion generation on provided frame information Table 1 shows the results of the automatic evaluation of the generated conclusions compared to their reference conclusions, for different variants of frame information provided as part of the input. We report results for three evaluation measures: ROUGE-1, ROUGE-L, and BERTscore (F1-score). As baseline, we rely on a model version that only relies on premise and topic as input, but does not include any frame information (*no frame*). We can observe that adding information about the frame in the three specifications *specF, genFArg, genFConcl* has a positive impact on the generated conclusions, increasing results between 1.1 and 4.5 points for ROUGE-1, between 1.1 and 3.9 points for ROUGE-L, and between 1.0 and 4.4 points for BERTscore. The single frame specification with the most signifi-

[6]The frame distribution of the test set is similar to the frame distribution of the selected samples, having most "other" (40%) and "economic" (17%) generic frames (argument).

Configuration	Rouge1	RougeL	F1-BERTs.
no frame	29.1	26.4	29.4
specF	+2.1	+1.8	+2.2
genFArg	+1.6	+1.2	+1.9
genFConcl	+2.5	+1.9	+1.5
genFArg+genFConcl	+1.1	+1.1	+1.1
specF+genFArg	+1.3	+1.1	+1.0
specF+genFConcl	+1.9	+1.5	+2.0
all 3 frames	**+4.5**	**+3.9**	**+4.4**

Table 1: Automatic scores for various frame configurations (issue-**specific** **f**rame, **gen**eric **f**rame from **arg**ument, **gen**eric **f**rame from **concl**usion) without informative label smoothing and conclusion reranking

Configuration	Val	Nov	Both	spec-f	gen-f
random	*0*	*7*	*0*	*10*	*33*
no frame	50	50	17	67	78
specF	67	37	10	**90**	**89**
genFArg	**73**	37	10	87	83
genFConcl	67	50	13	77	78
specF+genFArg	40	**63**	7	80	72
specF+genFConcl	60	47	**20**	77	78
all 3 frames	70	40	10	83	83
reference	*73*	*73*	*47*	*83*	*83*

Table 2: Manual evaluation study: ratio of conclusions fulfilling the criteria of Validity, Novelty, both validity and novelty, and relatedness to the target issue-specific frame and the generic frame (argument), based on the majority votes for various frame configurations (issue-**specific** **f**rame, **gen**eric frame from argument, generic frame from conclusion), random and human-written reference conclusions, in %.

cant impact is the generic frame (conclusion) (*genFConcl*) for ROUGE-1 and ROUGE-L and the issue-specific frame (*specF*) for BERTscore. Considering combinations of two frame specifications (*genFArg+genFConcl, specF+genFArg, genFConcl+specF*) yields worse results in all cases, compared to using a single source of information. However, using all three frame specifications yields the best result, with improvements of 4.5, 3.9, and 4.4 points for ROUGE-1, ROUGE-L, and BERTscore, respectively.

Table 2 shows the results of the manual evaluation, where three human raters decided whether the generated conclusion is i) valid, ii) novel, iii) directed towards the issue-specific frame and iv) directed towards the generic frame (argument). The table shows the percentage of conclusions for which the majority of annotators agree that the conclusion is valid/novel/directed towards the desired frame. Manual assessment of a *random* conclusion (sampled from all generated and reference conclusions) and of the *reference* conclusion provide a lower and upper bound for our approach. The results of the manual evaluation corroborate that each *single frame* configuration has a positive impact on validity between +17% to +23% points – at the expense of no improvement or even decrease in novelty. Providing frame information as input also yields an increase in frame-relatedness (up to +23% points). For combinations of two or more frame specifications we see a mixed pattern: a decrease in validity (−10% points) and increase in novelty (+13% points) for *specF+genFArg* and the reverse pattern for *specF+genFConcl* and *specF+genFArg+genFConcl* (+10%/+20% points regarding validity and −3%/−10% points regarding novelty). However, we observe that in view of generating both valid and novel conclusions, all configurations except for *specF+genFConcl* (+18%) perform below the unframed baseline. At the same time, all configurations clearly improve upon the baseline in generating a conclusion that fits the desired issue-specific frame (see Table 2), with improvements ranging from +10% (specF+genFConcl) to +23% (specF) points. Regarding the frame relatedness to the generic frame, we see clear improvements over the baseline for 3 out of 6 configurations, ranging from +5% (genFArg, specF+genFArg+genFConcl) to +11% (specF) points.

Below, we investigate the impact of further strategies on the four configurations that were rated as best with respect to a single dimension: *specF* (for specF + genFArg), *genFArg* (for validity), *specF+genFArg* (for novelty), and *specF+genFConcl* (for both validity and novelty).

Impact of strategies to boost frame-sensitive conclusion generation To measure the impact of our strategies for boosting frame-sensitive conclusion generation, the annotators were asked to rate the validity and novelty of the conclusions in a pairwise setting with and without activated label smoothing and/or conclusion reranking, and had to rate whether they found an increase, tie or decrease of novelty and/or validity.[7] Table 3 shows the results of this further manual evaluation, where next to *Val*, *Nov* and *Both* we show the absolute improvements in automatic BERTscore for each strategy.

[7]Our annotators evaluated up to 60 samples for conclusion reranking: the 30 arguments from the first annotation rounds + 30 new arguments for input variants: *no frame, specF+genFArg, specF+genFConcl*.

Configuration	+ inf. label smoothing				+ conclusion reranking							
					frame-insensitive				frame-sensitive			
	BERT	Val	Nov	Both	BERT	Val	Nov	Both	BERT	Val	Nov	Both
no frame	+4	+13	+23	+10	**+13**	+10	-13	-5	n/a	n/a	n/a	n/a
specF	+7	-13	**+23**	+7	+8	**+27**	-17	+3	+7	+9	-13	0
genFArg	+2	-33	+23	0	+8	**+27**	**-10**	0	+7	**+20**	-10	**+3**
specF+genFArg	+2	+3	-27	0	+10	+20	-20	-3	+8	+17	**-7**	0
specF+genFConcl	**+8**	**+30**	+13	**+13**	+11	+8	-17	-2	**+8**	+13	-12	-3

Table 3: Evaluation of additional strategies for boosting frame-sensitive conclusion generation automatically (F1-BERTscore) and manually (majority votes per conclusion in Validity, Novelty, Both) for various frame configurations (issue-**specific** **f**rame, **gen**eric frame from **arg**ument, **gen**eric frame from **concl**usion). The deltas measure the difference to the next lower model complexity (w/o any additional strategy/ only informative label smoothing) in %.

Informative label smoothing has a positive impact w.r.t. to the BERTscore (+2% to +8%)). With respect to validity, it improves results in 3 out of 5 configurations, while with respect to novelty, it improves results in 4 out of 5 configurations in the range of +13% to +23%. We thus see a clear positive impact on novelty.

Conclusion reranking improves the BERTscore in all configurations, both in the frame-insensitive (+8% to +13%) and the frame-sensitive variant (+7% to +8%). Both variants have a positive impact on validity, improving results between +8 and +27% (frame-insensitive variant) and between +9 and +20% (frame-sensitive variant). However, both variants do not consistently improve the number of conclusions that are regarded as both valid and novel across configurations, with differences ranging from -5% to +3%.

5.3 Discussion

Regarding the impact of conditioning conclusion generation by providing information about the desired frame, our results of both automatic and manual evaluations are generally positive. We see a clear improvement in the framing and the similarity of the generated conclusions to their reference conclusions. The results of our manual evaluation clearly point to a trade-off between *generating a valid vs. novel conclusion*, showing that it is very challenging to generate conclusions that fulfill both criteria (novelty and validity). For example, providing information targeting an issue-specific frame (*specF*) increases validity by 17% points while decreasing novelty by 13% points at the same time. There are other configurations, however, that resolve this trade-off better. The combination of issue-specific as well as conclusion-retrieved generic frames (*specF+genFConcl*) yields the best results in generating a conclusion that is both valid and novel (20%), outperforming the *no frame* baseline by 3% points. This configuration leverages information from the two different types of frames, providing the model information at different and thus complementary levels of granularity as proposed by Heinisch and Cimiano (2021).

Overall, the best configuration in terms of validity, judging from our manual majority votes, is the version that relies on issue-specific frame information as input in combination with informative label smoothing and frame-sensitive conclusion reranking (87%). Regarding novelty, the best configuration combines the issue-specific and generic frame (argument) with informative label smoothing but without conclusion reranking (67%). The configuration that excels in generating both novel and valid conclusions is the one that enriches the input with the issue-specific frame as well as the generic frame (conclusion), again using only informative label smoothing (40%).

In general, assessment by way of BERTscores does not correlate well with manual assessment of validity and novelty. While BERTscores improve in all cases when applying informative label smoothing and especially conclusion reranking (up to 37.6), the manual evaluation of validity and novelty in those configurations is quite mixed. Many configurations improve validity at the cost of novelty and the other way round. In general, informative label smoothing has a positive impact on novelty. It seems that preferring conclusions that include tokens that frequently occur with the desired frame is driving the model to leave its comfort zone and take risks in generating conclusions with novel elements with the downside that some of these conclusions seem not to be perceived as valid. In contrast, conclusion reranking, which

Wave power – premise:	
Wave power has the potential to provide 5-10% of US energy supply, according to the New York Times. *(issue-**spec**ific **f**rame: energy / **gen**eric **f**rame (**arg**ument): Capacity and resources)*	
reference	Wave power can contribute a significant amount of energy
no frame	Wave power can significantly increase energy supply
+smooth	Wave power has the potential to replace coal
genFArg	Wave energy has the potential to power the US
specF+ genFArg	Wave power is a major source of clean energy
+smooth	Wave power can supply a significant amount of energy
+concl. rerank	Wave power can provide 10% of US energy supply

Table 4: Case study showing the effects of different configurations, including informative label smoothing and frame-insensitive conclusion reranking

learned to optimize primary the BERTscore between premise and conclusion candidate, restricts novel content by selecting premise-similar content, ensuring validity at the expense of novelty.

Case study Table 4 shows clear differences in wording between the conclusions generated using different configurations for a hand-selected example. The conclusion generated without frame information mentions 'a significant increase of energy supply' vs. 'significant amount of energy' (reference conclusion). When informative label smoothing is active, the conclusion mentions the 'potential to replace coal', bringing in a novel element not mentioned in the premise. Adding the generic frame (Capacity and resources) interestingly leads to emphasizing the 'potential to power the US'. Adding information about the issue-specific frame (energy) changes this back to talking about wave energy as a 'major source of clean energy'. Conclusion reranking picks up specific elements of the premise e.g. ('10 % of US energy supply'), but lacks novel elements compared to the given premise. The case study clearly shows that we can control conclusion generation in ways intended by our methods. However, it also shows that the observed impacts are subtle.

6 Conclusion

In this paper we have studied the question of how to condition the generation of argumentative conclusions from premises using a transformer-based fine-tuning approach based on pre-trained language models. We have shown the positive impact of different strategies to bring the generated conclusions closer to the desired frame during inference while showing that proposing conclusions that are perceived as both valid and novel by humans is challenging, especially since these two dimensions seem to stand in a trade-off that renders their joint optimization difficult. Our results clearly show that the proposed strategies have the potential of improving either novelty or validity. In future work we aim to investigate the factors that contribute to validity and novelty in more detail. Especially we aim to understand how to control the trade-off between validity and novelty better to maximize the likelihood of generating conclusions that fulfill both criteria.

Acknowledgements

We are grateful to the anonymous reviewers for their valuable comments. This work has been funded by the DFG through the project ACCEPT as part of the Priority Program "Robust Argumentation Machines" (SPP1999).

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 2915–2925, Hong Kong, China. Association for Computational Linguistics.

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.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Maria Becker, Ioana Hulpus, Juri Opitz, Debjit Paul, Jonathan Kobbe, Heiner Stuckenschmidt, and Anette Frank. 2020. Explaining arguments with background knowledge. *Datenbank-Spektrum*, 20(2):131–141.

Maria Becker, Siting Liang, and Anette Frank. 2021. Reconstructing implicit knowledge with language models. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–24, Online. 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.

Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. Dataset independent baselines for relation prediction in argument mining. *Frontiers in Artificial Intelligence and Applications*, 326(Computational Models of Argument):45–52.

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

Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2019. Tacam: Topic and context aware argument mining. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 99–106, New York, NY, USA. Association for Computing Machinery.

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Wei Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *2021 International Conference on Learning Representations*.

Philipp Heinisch and Philipp Cimiano. 2021. A multi-task approach to argument frame classification at variable granularity levels. *it - Information Technology*, 63(1):59–72.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Xinyu Hua and Lu Wang. 2020. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria. INCOMA Ltd.

W Russell Neuman, Russell W Neuman, Marion R Just, and Ann N Crigler. 1992. *Common knowledge: News and the construction of political meaning*. University of Chicago Press.

Juri Opitz, Philipp Heinisch, Philipp Wiesenbach, Philipp Cimiano, and Anette Frank. 2021. Explainable unsupervised argument similarity rating with Abstract Meaning Representation and conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. 2020. Argumentative relation classification with background knowledge. In *Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 319–330. IOS Press.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Pavithra Rajendran, Danushka Bollegala, and Simon Parsons. 2016. Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 31–39, Berlin, Germany. Association for Computational Linguistics.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for 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.

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.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9048–9056.

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

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017b. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain. Association for Computational Linguistics.

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.

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*.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Appendix

A.1 Transformer-based generic frame classification model

To induce a generic frame classifier, similar to Heinisch and Cimiano (2021) we fine-tuned a RoBERTa-base (Zhuang et al., 2021) language model on the training section of the Media-Frames dataset (Card et al., 2015) (Section 3.1). We only considered text spans annotated by at least two annotators agreeing on the same frame. We encode the annotated text spans with RoBERTa-base and use the `[CLS]` head to predict a probability distribution over the 15 frame classes. The trained model obtained an accuracy of 58% on our held-out test split from the Media-Frames dataset.

A.2 Reference-less scores for our conclusion reranking approach

Below we describe a selection of different reference-less scores, which we use to help the conclusion reranker select an appropriate conclusion among several conclusion candidates. A first group of scores rate the conclusion candidate stand-alone, others measure the relation between premise and conclusion candidate, and some rely on the given frame information to rate the quality of a conclusion.

Conclusion-candidate-based scores To rate the quality of a conclusion candidate stand-alone, we measure its absolute token length as well as the number of non-stopword-tokens it contains and the ratio of non-stopwords-tokens to stopword-tokens. Further, we check for patterns that are typical for conclusions, such as *is, better than, should, therefore*.

Premise-&-conclusion-candidate-based scores Another way to rate the quality of a conclusion candidate is to characterize its relation to the premise. Here we take into account the relative conclusion candidate token length compared to the premise token length. Further, we measure coherence and grammaticality with the GRUEN-score (Zhu and Bhat, 2020) by concatenating the premise with the conclusion candidate. We also measure the similarity of the conclusion candidate and the premise using BERTscore[8] (Zhang et al., 2020), using the outputted precision, recall, and the F1-score.

Finally, we aim to assess the argumentative relation of the conclusion candidate and the premise, by building a model that classifies this relation into attack, no relation or support. To this end, we fine-tuned a DeBERTa-base language model (He et al., 2021) for natural language inference (NLI) classification. We build `SEP`-structured inputs consisting of topic, premise and conclusion candidates, using the argument dataset by Ajjour et al. (2019) with the same train and validation split as used in our presented evaluation (Section 5). From this argument dataset we obtain positive samples (entailment/ support class) by concatenating premises with their reference conclusion. To generate samples with the neutral target class "no relation", we provide premises with conclusions sampled from other dataset topics. To provide negative samples (contradiction / attack), we join each premise with a sampled conclusion from the same topic fulfilling the same issue-specific frame but having the opposite stance towards the topic. This information is provided by the dataset. In this way, we generated a balanced dataset from Ajjour et al. (2019)'s dataset. The model trained on this data reaches an accuracy of 86% on the test split (including the Ω-optimization split). Using this fine-tuned language model, we tag each conclusion candidate with the predicted entailment class

[8] using the 18th layer of `microsoft/deberta-xlarge` without an idf-weighting

probability and a score $P(Entailment) \cdot (1 - P(Contradiction)^2)$ to measure the risk of having a conclusion candidate which is attacked by its premise and therefore, not a valid conclusion.

Frame-sensitive scores If the input provides the issue-specific frame, we score the probability of the conclusion candidate belonging to this frame using a RoBERTa-base (Zhuang et al., 2021) language model with `[CLS] conclusion candidate [SEP] issue-specific frame label [SEP]` as input. The predicted value between 0 (not frame-related) and 1 (frame-related) is considered for the conclusion candidate selection. For fine-tuning such a model, we use the same train and development splits from the argument dataset of Ajjour et al. (2019) as above. We model the task as a regression task, assigning 1 for the edge case of a true issue-specific frame label and 0 for the edge case of a completely unrelated issue-specific frame label. For each positive sample that combines a conclusion with its ground truth frame label $refF$, we generate a negative sample by combining the conclusion with a frame label having the largest Word-Movers-Distance WMD (Kusner et al., 2015) $negRefFf$ to the reference frame label. To have a more fine-grained regression objective, we create additionally mixed samples by randomly sampling a frame label $randF$, having a ground-truth-score of $1 - \frac{WMD(refF,randF)}{WMD(refF,negRefF)}$. The resulting mean absolute error is 0.11 on the test split (including the Ω-optimization split).

In cases where the input contains generic Media-Frames information, we take into account the probability of matching that frame, using again a fine-tuned RoBERTa-model. We use the mode described in A.1.

A.3 Further insights into the manual study

We performed an extensive manual annotation study to assess the quality of the generated conclusions for the various settings.

A.3.1 Annotators and agreement

Our aim was to collect high-quality annotations. To this end, we accepted only paid students with higher education entrance qualification working on research projects related to argument mining. After qualifying questions related to the annotation study, including a positive and negative annotation example, each student annotated independently from the other.

Each sample was annotated three times. We split our annotation study into three rounds. The first round aims to find the best input frame configuration. In the second round, we explore the informative label smoothing. The third round rates conclusions generated using the conclusion reranking technique. The same 30 samples from the test split were used for all rounds, and all were evaluated by the same three annotators. In addition, the third round included a second bulk of 30 arguments to increase the statistic relevance. We invited two additional annotators to annotate the second bulk (keeping 1 of the previous annotators). Hence, five annotators participated in the annotation study in total.

The Fleiss-kappa-inter-annotator-agreements for the absolute judgments (yes/no) are 0.53 for validity, 0.22 for novelty and 0.4 for framing-relatedness. Among the absolute judgments, 6%, 4%, and 4% were undecided ("I don't know") for validity, novelty, and framing-relatedness, respectively. The moderate agreement for validity is relatively high for such an argumentative task. However, in general, the agreement on similar tasks has been shown to be quite low because of subjectivity (Gurcke et al., 2021). One source of this subjectivity is in the decision of where to draw the line between two categories (e.g., novel vs. not-novel).

The Fleiss-kappa-inter-annotator-agreements for the pairwise setting (Conclusion 1 is better/ equal/ Conclusion 2 is better) are 0.48 for validity, 0.36 for novelty and 0.41 for framing-relatedness.

A.3.2 Annotation interface

To give further insight about the annotation task and provided instructions, Figure 4 shows a screenshot of the annotation interface in the pairwise setting, using a dummy sample. The different formatting styles (colors, borders and font style) of the conclusion boxes result from the selected rating. These interactive styles support the annotators by visualize their rating.

A.4 Analysing the gap between higher BERTscores and lower manual ratings

To provide a better understanding of the discrepancy between BERTscores and manual ratings especially in the case of activated conclusion reranking (in combination with informative label smoothing), we list a few samples in 5.

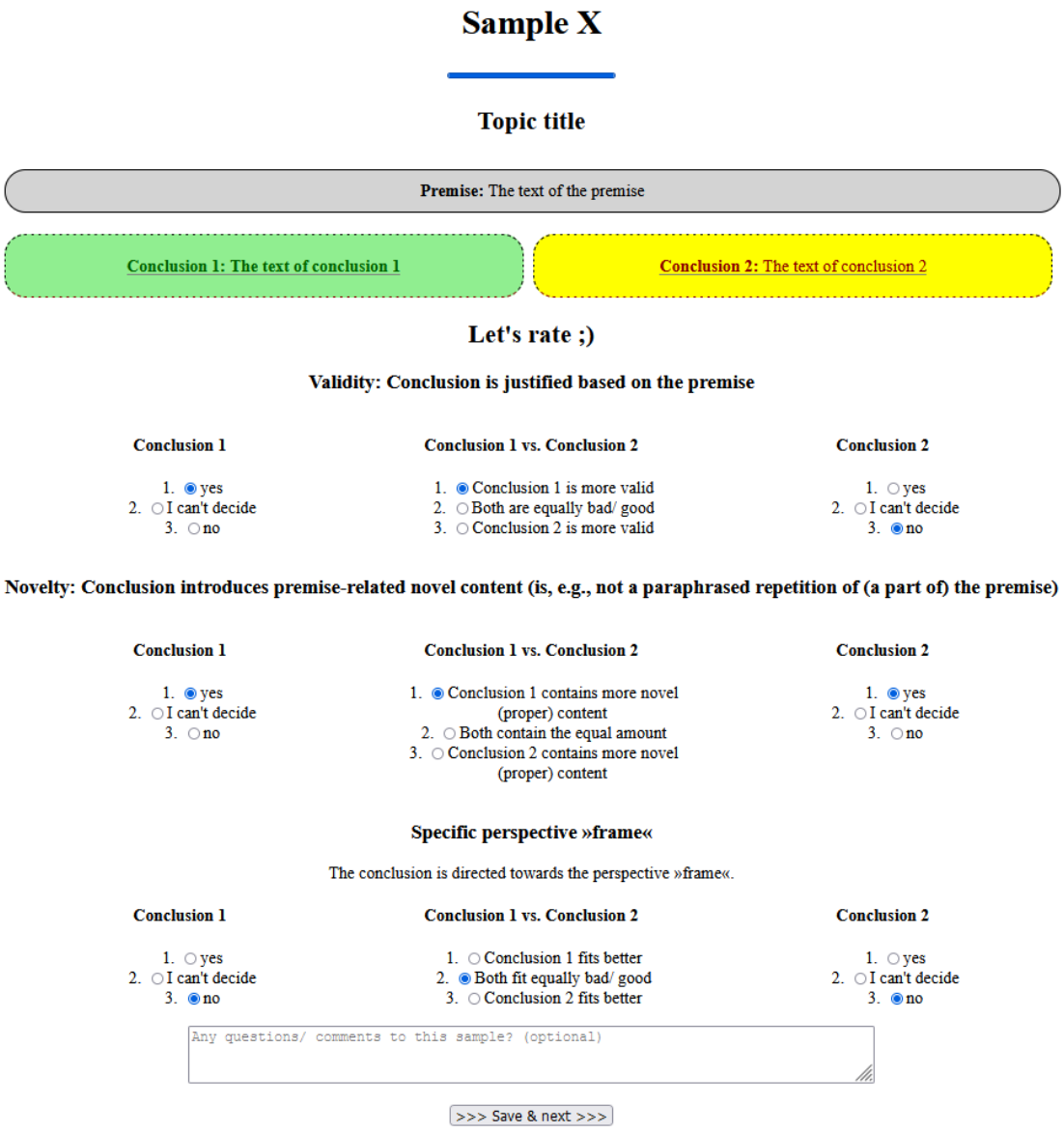

Figure 4: A screenshot of the used annotation tool

Topic+*issue-specific frame*+Premise ↦ *reference conclusion*	**w/o conclusion reranking**	**conclusion rerank**	**rating for concl. rerank**
Withdrawing from Iraq *political progress* Although the constitution and formation of a legitimate government are major achievements, there is still much work to do. Sunni Arabs have to be convinced that the new settlement is in their interests and be drawn into government. This requires that they, and others, are convinced that the Iraqi government will stand the test of time, which requires the support of the US both politically and militarily ↦ *Political progress has been made but must be continued with coalition support.*	Sunni Arabs have to be convinced that the Iraqi government is legitimate	Iraqi Sunni Arabs must be convinced that a legitimate government exists.	validity: - -, novelty: - -, issue-specific-frame-relatedness: -
Withdrawing from Iraq *soldiers* It is odd to argue that a rapid withdraw would be risky to the lives of troops. Keeping them in Iraq is obviously risky, and particularly if it is extended over many years. ↦ *Keeping troops in Iraq is obviously more dangerous than a rapid withdrawal.*	Withdrawing from Iraq would not risk the lives of troops	Withdrawing from Iraq would risk the lives of troops	validity: - - -, novelty: -, issue-specific-frame-relatedness: -
Withdrawing from Iraq *iraqi will* Sheikh Mahmood Ejemi, head of the Ejmani tribe in Hiyt, believes the improved security could quickly unravel without sufficient US support. - ””I advise the Americans to withdraw only when Iraqis can secure and achieve security and have a strong and capable military force to protect the borders and the populace. We need to have a national government that protects Iraqis, not works to isolate and kill them, like it is doing now. We need US support in fighting sectarian militias and al-Qaeda” ↦ *Many Iraqis support the continued presence of US forces*	The US should not withdraw before the Iraqi government is sufficiently secure	Iraqis will not accept US withdrawal until they have a strong military.	validity: - - -, novelty: - -, issue-specific-frame-relatedness: + +
Vegetarianism *health* Almost all dangerous types of food-poisoning (e.g. E-coli, salmonella) are passed on through meat or eggs. Close contact between humans and animals also leads to zoonosis – diseases such as bird ‘flu which can be passed on from animals to humans. Hunters eating apes and monkeys is thought to have brought HIV/AIDS to humans. And using animal brains in the processed feed for livestock led to BSE in cattle and to CJD in humans who ate beef” ↦ *Meat-eating is linked to a range of serious illness such as food-poisoning.*	Vegetarians are not immune to diseases of animals	Vegetarians are vulnerable to food poisoning.	validity: -, novelty: - -, frame-relatedness: -
Video surveillance *privacy*: It is certainly not the case that people monitor all security cameras closely 24/7. Most surveillance tapes are rarely seen. Usually surveillance cameras are only viewed if they have filmed a crime and are viewed only to catch criminals, not to invade people’s privacy or stalk people.” ↦ *Surveillance cameras are not closely monitored and are only usually viewed if a crime has taken place.*	Privacy infringements are rare with surveillance cameras	Surveillance cameras are rarely viewed to catch criminals.	validity: - - -, novelty: - - -, issue-specific-frame-relatedness: - -

Table 5: Examples of generated conclusions in which the frame-insensitive conclusion reranking technique clearly leads to better BERTscores (covering more parts of the reference conclusion) than the conclusion without reranking but receiving worse scores in the manual evaluation. Each − reflects a dispreference to the conclusion-reranking-output, while each + represents a preference rating.