MULTIOPED: A Corpus of Multi-Perspective News Editorials

Siyi Liu **Sihao Chen** **Xander Uyttendaele** **Dan Roth**
University of Pennsylvania
{siyiliu, sihaoc, xanderu, danroth}@seas.upenn.edu

Abstract

We propose MULTIOPED[1], an open-domain news editorial corpus that supports various tasks pertaining to the argumentation structure in news editorials, focusing on *automatic perspective discovery*. News editorial is a genre of persuasive text, where the argumentation structure is usually *implicit*. However, the arguments presented in an editorial typically center around a concise, focused thesis, which we refer to as their *perspective*. MULTIOPED aims at supporting the study of multiple tasks relevant to *automatic perspective discovery*, where a system is expected to produce a single-sentence thesis statement summarizing the arguments presented. We argue that identifying and abstracting such natural language *perspectives* from editorials is a crucial step toward studying the implicit argumentation structure in news editorials.

We first discuss the challenges and define a few conceptual tasks towards our goal. To demonstrate the utility of MULTIOPED and the induced tasks, we study the problem of perspective summarization in a multi-task learning setting, as a case study. We show that, with the induced tasks as auxiliary tasks, we can improve the quality of the perspective summary generated. We hope that MULTIOPED will be a useful resource for future studies on argumentation in the news editorial domain.

1 Introduction

News editorial is a form of persuasive text that conveys consensus opinion on a *controversial topic* from the editors of a newspaper. Much like an argumentative essay, a news editorial centers around a thesis, which represents the authors' *perspective* on the topic. Usually, a news editorial argues in favor of the authors' *stance* on the topic, and is substantiated by extensive factual *evidence*.

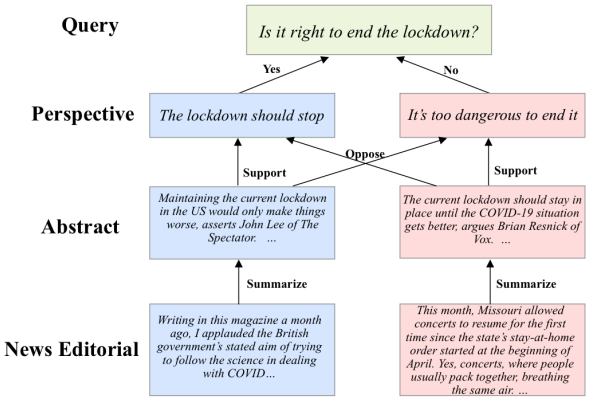

Figure 1: Structure of MULTIOPED. For each *query* on a controversial topic, two (rather long) news editorials respond to the query from different point-of-views. Each editorial comes with a single paragraph *abstract* plus a one-sentence *perspective*, that abstractively summarizes the editorial's key argument in the context of the query. The two resulting perspectives serve as responses with opposite *stance* to the query.

As news editorials function as professionally produced written discourse for conveying media attitude and guidance, they have traditionally been studied by the community as a rich resource for many argumentation-related tasks. (Wilson and Wiebe, 2003; Yu and Hatzivassiloglou, 2003; Bal and Saint-Dizier, 2009).

This work targets the problem of developing computational methods to identify and comparatively analyze the authors' *perspectives* and supporting arguments behind news editorials. One challenge to studying the argumentation structure in news editorials is that its elements are rarely expressed explicitly (El Baff et al., 2018). For example, Figure 1 shows two news editorials holding opposite views on whether a lockdown should continue. However, neither of them present their key perspectives explicitly. Instead, the *perspective* is conveyed through subtle rhetoric strategies to either affirm or challenge the readers' stance from

[1]The authors would like to thank Daniel Ravner, the CEO of www.theperspective.com, for granting access to data from the site for academic research.

prior belief on the topic, as a study by El Baff et al. (2018) discovers. As Figure 1 shows, the statement "*The lock down should stop*" concisely summarizes the perspective expressed in the article on the left. We refer to such statements as "*perspectives*" throughout the paper. The ability to abstractively summarize the perspectives from the editorial would allow us to understand multiple topic-aligned editorials in context and reason about their inter-editorial argumentation structure.

To facilitate research along the line, we collect data from THEPERSPECTIVE[2] website, and construct MULTIOPED, an open-domain English news editorial corpus that supports various tasks pertaining to the argumentation structure in news editorials, focusing on *automatic perspective discovery* (Chen et al., 2019). The structure of the data is shown in Figure 1. For each of the 1,397 natural language *query* on a different topic in our dataset, it features two (rather long) news editorials. Each editorial features a single-sentence perspective, which is abstractively summarized from the editorial by human experts. A short *abstract* then highlights the details in the editorial that support the *perspective*. The *perspectives* of the two editorials represents responses of opposite *stances* towards the query.

Naturally, the structure of the dataset induces a range of important argumentation-related natural language understanding tasks. For instance, the presence of the summary *perspective* allows for stance classification (Hasan and Ng, 2013) with respect to the query, which arguably is more tangible than inferring the stance from the entire editorial. Another example task is the conditional generation of the *perspective* from the abstract/editorial, which relates to the widely studied task of *argument generation* (Hua and Wang, 2018; Alshomary et al., 2020). We defer the more detailed description of the induced tasks to Section 3.

One key advantage of MULTIOPED that is absent from earlier datasets is that a large number of argumentation-related tasks can be studied jointly using a single high quality corpus. To demonstrate this benefit and the utility of the MULTIOPED dataset [3] along with its induced tasks, we study the problem of perspective summarization in a multi-task learning setting. We employ perspective relevance and stance classifications as two auxilliary tasks to the summarization objective. Our empirical and human analysis on the generated summaries show that the multi-task learning setting improves the generated perspectives in terms of the argument quality and stance consistency.

In summary, our contributions in this work are three-fold. First, we propose a conceptual framework for identifying and abstracting the *perspectives* and the corresponding argumentation structure in news editorials, and define a set of tasks necessary for achieving this goal. Second, we propose the MULTIOPED dataset, a news editorial dataset that induces multiple argumentation-related tasks. Third, we demonstrate the utility of our multi-purpose dataset and induced tasks, by using the perspective summarization task as a case study. We include the induced tasks as auxiliary objectives in multi-task learning setting, and demonstrate their effectiveness to perspective summarization.

2 Design Principles

Our goal of perspective discovery follows similar definition proposed by Chen et al. (2019), and is closely related to a widely studied area of argumentation mining, i.e. identifying the argumentation structure within persuasive text (Stab and Gurevych, 2014b; Kiesel et al., 2015). However, most studies in this domain focus on extractive methods, which becomes less applicable to our study. As the arguments are usually presented in an subtle and implicit way in news editorials, we instead focus on the generation methods for the perspectives. This closely resembles the argument conclusion generation task (Alshomary et al., 2020). One key distinction here is the presense of *query* to provide topic guidance during the perspective generation.

Compared to other conditional text generation tasks, perspective generation subjects to a few more constraints with respect to the argumentation structure. For example, the perspective must constitute the same stance (Hasan and Ng, 2013) as the editorial towards the query. On the other hand, while the editorial may cover content not directly related to the query, the generated perspective must present a relevant argument in the query's context. Such structural constraints can be studied in the format of classification problems. And being able to study such problems along side the perspective summarization task on one high-quality corpus is important in our case, as it opens up the probability of

[2] https://www.theperspective.com/perspectives/

[3] Our code and data is available at http://cogcomp.org/page/publication_view/935

Dataset	Source	Open Domain	Cross Article	Abstractive
AraucariaDB (Reed et al., 2008)	News Ed.	✓	×	×
(Stab and Gurevych, 2014a)	Essay	✓	×	×
(Eckle-Kohler et al., 2015)	News	✓	×	×
(Hua and Wang, 2018)	Reddit/Wiki.	*Politics* only	×	✓
PERSPECTRUM (Chen et al., 2019)	Debate	✓	×	✓
MULTIOPED	News Ed.	✓	✓	✓

Table 1: A comparison across datasets with similar purpose to MULTIOPED. We compare the datasets along three dimensions. *Open Domain*: whether the dataset features a wide variety of topics. *Cross Article*: whether the argumentation structure between documents are annotated. *Abstractive*: whether the elements in argumentation structure is abstractive or extractive.

modeling the tasks jointly. We show the benefit of doing so by presenting a case study in section 5.

As the query provides topic guidance, it allows for the study of the topic-aligned pairs of editorials which presents counter-arguments to each other. Such property is absent from notable datasets of similar purposes to ours, as shown in Table 1. AraucariaDB (Reed et al., 2008) is the first effort to provide large-scale annotations of dense argumentation structure within individual news editorials. Stab and Gurevych (2014a); Eckle-Kohler et al. (2015) provide resources for extractive argumentation structure in persuasive essays and news articles, respectively. Later works (Hua and Wang, 2018; Chen et al., 2019) focus on the abstractive generation or identification of arguments from web corpora. All of these datasets focus on studies of argumentation structure within individual document. Instead, our proposed dataset presents the opportunity to study the cross-document argumentation structure.

Instance	Size	Avg. Len.	Min	Max
Query	1,397	7.4	3	15
Perspective	2,794	6.1	2	10
Abstract	2,794	101.9	47	160
Article	2,584	918.6	74	7,608

Table 2: Statistics of the MULTIOPED dataset. Size represents the number of each valid instance, Avg. Len. indicates the average length of each instance in terms of the number of tokens split by space, and Min and Max represent the number of tokens of the shortest and longest texts of each instance.

3 MULTIOPED and Induced Tasks

Following the design principles outlined in the previous section, we propose a topic-aligned English news editorial corpus, MULTIOPED. The structure of an example instance in MULTIOPED is shown in Figure 1. To clarify our description of the dataset, we use the following notations. Let q be a query about a controversial topic. Each q in the dataset is paired with two editorials e_{pro} and e_{con}, that constitute *supporting* and *opposing* stances to the query q respectively. Each editorial is abstracted into and a single-sentence *perspective* p, which provides a high-level summarization of the key argument presented in the editorial. The premises, or relevant details to support the perspective, forms the abstract a.

Naturally, the relation between these elements induces several tasks, most of which encompass similar definitions to existing argumentation-related tasks. We define and describe the tasks and their connection to our end goal of perspective discovery below.

1. *Generating an Abstract*: Given an editorial e, a system is expected to identify and summarize the relevant arguments into an abstract paragraph a to the context provided by the query q. This is closely related to the task of argument synthesis (El Baff et al., 2019; Hua et al., 2019). We set aside this problem in our case study in section 5, and use the abstract provided by the dataset.

2. *Perspective Summarization*: Given the generated abstract a and the query q, a system is expected to generate the perspective p, a concise summary of the arguments presented in a. Conceptually, this problem resembles the task of argument conclusion generation (Alshomary et al., 2020). We adopt a slightly different setting where the target topic is expressed in the form of a natural language query.

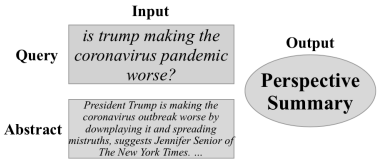

Figure 2: Perspective Summarization: Generate an single sentence argument that represents the key *perspective* expressed in the news editorial.

3. *Stance Classification*: Our goal is to infer the editorials e's stance towards a query q. The generated perspective p from editorial e allows us to focus on a simpler task definition of classifying the stance of the perspective to the query q (Hasan and Ng, 2013; Bar-Haim et al., 2017).

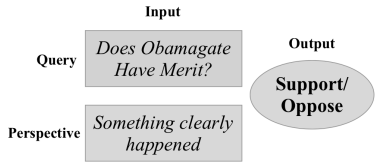

Figure 3: Stance Classification: Decide if the perspective supports or opposes the query.

4. *Assessing the Relevance of Perspective*: We want to measure the validity of the perspective by assessing whether the perspective presents a relevant argument towards the query (Chen et al., 2019; Ein-Dor et al., 2020). This can be formulated as a classification problem with the query q and a perspective p as inputs, as we show in section 5.

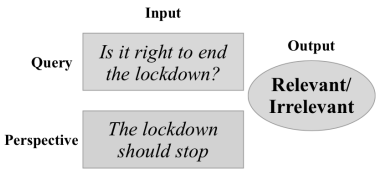

Figure 4: Relevance classification: Decide if the perspective is relevant to the query.

4 Dataset Construction

4.1 Data Collection

We extract the query, editorial article pairs, abstract paragraph pairs, along with their perspective summaries from THEPERSPECTIVE[4] website. The website presents controversial topics in the form of queries. For each query, two related editorial articles with opposing views from different sources are selected by the writers from the website. The writers create a concise one-sentence summary of each article as the response to the query, and an abstract paragraph to summarize the relevant arguments from the article. An example structure of the data is shown in Figure.1.

We use BEAUTIFULSOUP [5] and NEWSPAPER3K [6] to extract and clean the perspective and news data.

4.2 Crowdsource Verification & Annotation

To verify the structure from the website, and collect additional annotations, such as stance of the perspectives, we conduct a few annotation experiments with Amazon Mechanical Turk[7]. For all of our annotation experiments, we require the workers to be located in the United States, as the controversial topics covered by the website are most applicable in the U.S. context. We also require the workers to have masters qualifications (i.e. Top performers recognized by MTurk among all workers). We compensate the workers $\$0.75, \$1.00$ and $\$1.25$ per 10 queries for the implicit reference resolution, topic annotation, and stance annotation tasks respectively. The compensation rates are determined by estimating the average completion time for each annotation experiments. Example screenshots of our annotation interface and more detailed annotation guidelines can be found in Appendix B.

4.2.1 Stance Annotation

In our dataset, each *query* is presented with two *perspectives* with opposite stance to the query. However, the raw data that we collected does not specify the stance of each perspective individually.

We ask two expert annotators label whether each perspective is offering a supporting or opposing view with respect to its query. The two experts discuss and adjudicate their decisions. We then ask on average three crowdsource workers per instance to verify the annotations.

From the annotations collected by experts, we find that 30 out of $1,397$ queries do not constitute a clear stance. Such queries are typically "open-ended" questions which cannot be responded with a `yes` or `no` answer, i.e. `why` or `what` questions. We leave these instances unlabeled and exclude them from the next verification step.

[4] https://www.theperspective.com/perspectives/

[5] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[6] https://newspaper.readthedocs.io/

[7] https://www.mturk.com/

To assess the quality of stance labels created, we randomly sample 500 perspectives, and ask three MTurk workers per instance to verify stance labels. We computed the inter-rater agreement fleiss' $\kappa = 0.81$ among workers, and the agreement between majority decision from works and the expert's adjudicated annotations is cohen's $\kappa = 0.92$. We describe how we measure the two types of agreements respectively in Appendix A.3.

4.2.2 Implicit Reference Resolution in Perspectives

Some of the perspectives in our dataset have implicit references to certain subjects in the query. For instance, for a query *"Is Trump Right To Criticize Mail-In Voting?"*, and a perspective *"It's far too risky for an election"*, the word "It" in the perspective refers to "Mail-in Voting" in the query. As we assume that a perspective should presents a complete, valid argument on itself, we decide to replace such implicit reference in a perspective with the correct referent in the query. For example, the corrected perspective in the previous example would become *"Mail-in voting is far too risky for an election"*.

We ask one expert annotator to identify implicit references and make modifications for every perspective in the dataset. In total, $1,301$ out of $2,794$ perspectives are identified and corrected by the expert annotator. We ask three Turkers to verify that the modifications do not introduce any grammatical error or change the original meaning. We randomly sample 500 modified perspectives and present Turkers with the question of "Will this modification change the original meaning or introduce grammar error?". The percentage of majority answers being "No" is 84%. We include both changed and original versions of the perspectives in our datasets.

4.2.3 Topic Annotations

We create 9 topic labels according to the categorization from THEPERSPECTIVE website and major news outlets. We then as ask three MTurk workers to assign one of the 9 topic labels to each query. We regard the majority answer by the Turkers as the annotation for its topic category. In cases where all three annotators choose different categories (43 cases out of all 1397 queries), we label it as `other topics`. We show the distribution of topic categories in Figure 5. The inter-agreement among three annotators for this 9-class classification task is $\kappa = 0.65$

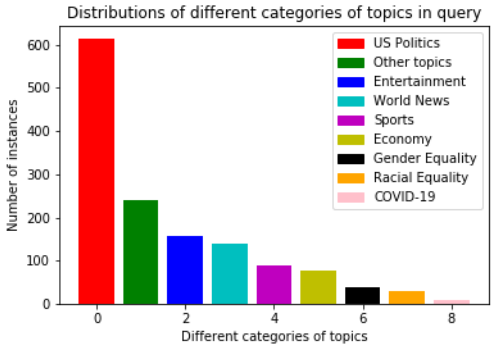

Figure 5: Topic distribution of the 1397 *queries* in MULTIOPED. Note that the two editorials for each query fall in the same topic category as the query.

4.3 Dataset Statistics

MULTIOPED consists of 1,397 queries about different news topics. Each query is presented with two perspectives, two abstracts and two linked news editorials. Despite a few stale urls and invalid redirections, we manage to extract the text for 2,584 news editorials. More detailed statistics are reported in Table 2.

5 Case Study: Multi-task Learning for Perspective Summarization

5.1 Multi-Task Framework

To demonstrate the benefits of modeling the induced tasks on the argumentation structure, we present a case study on the task of perspective summarization. Given a query and an abstract from the related editorial, a system is expected to produce a concise and fluent summary perspective for the editorial. In addition, the generated perspective ideally should satisfy a few structural constraints with respect to the query. For instance, the generated perspective must constitute the same stance as the editorial towards the query. Also the perspective should be relevant in the context of the query. The two requirements resemble the "stance classification" and "perspective validity" tasks defined in section 3 respectively.

Motivated by this, we study the two tasks together with perspective summarization in a multi-task learning framework. We choose BART (Lewis et al., 2020) as our base summarization model. BART is a pretrained auto-regressive transformer (Vaswani et al., 2017) encoder-decoder model, that have been proven effective in conditional text generation and other NLP tasks.

Model	ROUGE$_1$	ROUGE$_2$	ROUGE$_L$	BERTSCORE	REL. %	STANCE %
BART	28.24	11.34	26.96	88.67	91.91	72.32
+ *Rel*	28.35	11.51	27.12	88.69	92.98	72.68
+ *Stance*	28.19	11.53	26.93	**88.75**	91.25	73.39
+ *Rel* & *Stance*	**29.18**	**11.92**	**27.94**	88.74	**94.64**	**74.29**

Table 3: Results of our multitask *perspective* summarization models. We compare to BART as a baseline, and experiment with different combinations of the auxiliary tasks. We report the F_1 score under ROUGE$_{\{1,2,L\}}$ and BERTSCORE metrics, as well as the percentage of summaries with the correct relevance and stance label, as predicted by our pretrained classification models respectively. See Appendix A for training details and hyperparameters settings.

Model	RANK	%1ST	REL.	STANCE
BART	2.09	49.50	77.00	70.50
+ *Rel*	**1.74**	**60.50**	**87.00**	70.50
+ *Stance*	1.78	58.50	83.00	**79.50**
+ *R* & *S*	1.76	59.00	82.00	69.00

Table 4: Human Evaluations results. "RANK" shows a model's averaged rank judged by the raters ($1 =$ `best`, $4 =$`worst`) "%1ST" represents the percentage of generated summaries from one model that are ranked the best. We allow ties in the ranking. REL. and STANCE are the percentages of generated summaries that are relevant to and have the correct stance with respect to the query.

We start with a pretrained BART base model with $139M$ parameters, and finetune the model to output the target perspective given the query and abstract concatenated as input. In addition, we put two separate linear layers over the pooled embeddings of the last decoder layer, and produce two binary labels indicating the relevance and stance correctness of the generated summary respectively. The two tasks and the perspective summarization are learned jointly, and share the underlying parameters from BART.

One obvious challenge in the setup is that we do not have access to the ground truth stance and relevance label for the generated summaries during training. To address this, we adopt similar strategies as in knowledge distillation (Hinton et al., 2015). We first train two separate BERT (Devlin et al., 2019) models for stance and relevance classificaiton respectively. Due to the size limit of our dataset, we pretrain both models on the PERSPECTRUM dataset (Chen et al., 2019), which contain 7,000 instances of training data, with similar formats and definition to our $(query, perspective)$ pairs. We further fine-tune the models on our training set. When measured against our test set, the relevance and stance models achieve binary accuracy of 92% and 75% respectively.

During the perspective summarization model training, we use the pre-trained BERT models for relevance and stance classification to predict labels for each generated summary. We expect the BART plus linear layers to "mimic" the predictions made by the two pretrained BERT models respectively. Specifically:

$$\mathcal{H}_{\text{Q}} = \text{EOS}(\mathcal{D}_{\text{BART}}(\mathcal{E}_{\text{BART}}(\mathcal{Q})))$$
$$\mathcal{H}_{\text{A}} = \text{EOS}(\mathcal{D}_{\text{BART}}(\mathcal{E}_{\text{BART}}(\mathcal{A})))$$

We feed the query and the abstract separately through the BART encoder ($\mathcal{E}_{\text{BART}}$) and decoder ($\mathcal{D}_{\text{BART}}$). We get their hidden representations $\mathcal{H}_{\text{Q}}$ and $\mathcal{H}_{\text{A}}$ as the embedding of the end-of-sentence ('</s>') token from the decoder. We then concatenate $\mathcal{H}_{\text{Q}}$ and $\mathcal{H}_{\text{A}}$, and feed the concatenation to the two linear layers. Finally, a softmax layer is applied to get stance/relevance predictions $\tilde{y}_{rel}$ and $\tilde{y}_{stance}$.

$$\tilde{y}_{stance} = \text{SOFTMAX}(W_s^T[\mathcal{H}_{\text{Q}}, \mathcal{H}_{\text{A}}])$$
$$\tilde{y}_{rel} = \text{SOFTMAX}(W_r^T[\mathcal{H}_{\text{Q}}, \mathcal{H}_{\text{A}}])$$

Next, We feed the query and the generated summary to the two pretrained BERT classification models to get the soft stance and relevance labels y_{rel} and y_{stance}. We use two mean square error (MSE) loss terms to measure the discrepancy between the BART predictions and the soft labels.

$$\mathcal{L}_{\text{REL}} = \text{MSELOSS}(y_{rel}, \tilde{y}_{rel})$$
$$\mathcal{L}_{\text{STANCE}} = \text{MSELOSS}(y_{stance}, \tilde{y}_{stance})$$

We combine $\mathcal{L}_{\text{REL}}$ and $\mathcal{L}_{\text{STANCE}}$ with the summarization objective, $\mathcal{L}_{\text{SUM}}$, which is the negative log-likelihood loss between generated and target perspective. The auxiliary losses $\mathcal{L}_{\text{REL}}$ and $\mathcal{L}_{\text{STANCE}}$

are weighted by tunable hyperparameters α_1 and α_2 respectively.

$$\mathcal{L} = \mathcal{L}_{\text{SUM}} + \alpha_1 \cdot \mathcal{L}_{\text{REL}} + \alpha_2 \cdot \mathcal{L}_{\text{STANCE}}$$

5.2 Results

5.2.1 Automatic Evaluations

Table 3 shows our evaluation results of our multi-task model with different combinations of auxiliary tasks. The reported results are averaged over three trained models with different random initializations. We first evaluate the generated perspective summaries against the target perspective with ROUGE (Lin, 2004) and BERTSCORE (Zhang et al., 2020) metrics. We observe that relevance and stance auxiliary tasks both increase the ROUGE and BERTSCORE, and combining the two objectives yields the best performing model under the ROUGE metrics.

To empirically verify whether the perspectives generated by our multi-task model are improved in terms of the relevance and stance correctness, we again use the two pretrained BERT classifiers to measure the percentage of generated summary with correct relevance and stance label. The results potentially suggest that by "mimicing" the predictions made by the two pretrained classifiers, our multi-task framework is able to generate summaries with higher quality along the two dimensions.

5.3 Human Evaluations

We randomly sampled 100 instances of abstracts with query from the test set, and ask two human raters to judge the quality of perspectives generated by the four systems. For the four summaries generated from an abstract by the different systems, we shuffle their order and ask the raters to rank each summary by the overall quality, with four criteria considered (1) *Fluency* (2) *Grammatical Correctness* (3) *Faithfulness to the arguments offered in the original abstract* (4) *Salience*. We allow ties among different summaries. We report their averaged ranks and the number of times a system is ranked first place in Table 4. The results are the averaged scores between the two annotators, and the level of agreement between them for this 4-class ranking task is $\kappa = 0.35$.

For each summary, we ask the raters to annotate whether it (1) represents a relevant argument to the query (2) constitutes the correct stance as the target stance label. The kappa agreement between the two raters for these two tasks are 0.54 and 0.70, respectively. We show the human evaluation results in Table 4. We observe that while both the relevance and stance auxiliary tasks improve the quality of the generated perspective, combining the two auxiliary tasks does not guarantee a better summary quality.

Query	Should trump accept democrats' gov't spending bill?
BART	A shutdown is the best deal he can get.
+ REL	Trump should accept the gop budget.
Gold	This deal is the most achievable compromise.

Table 5: An example where relevance auxiliary task helps the perspective summarization process

Query	Is apple's iphone x technology any good?
BART	Apple's new iPhone X offers many great opportunities
+ STANCE	Apple's new face-recognition technology raises many ethical issues
Gold	Apple's new Iphone X raises many security concerns

Table 6: An example where stance auxiliary task helps the perspective summarization process

5.4 Analysis and Discussion

The results on ROUGE, BERTSCORE and human evaluation suggest that the perspective summarization model learning benefits from both the relevance and stance tasks. However, we also observe that the vanilla BART present a strong baseline in both automatic and human evaluations.

We list two typical cases where we observe the relevance and stance objectives improve the quality of the generated summary. For the query shown in Table 5, the BART model generates an out-of-context word "shutdown", which exists in the abstract, but is not applicable in the context provided by the query. The model with relevance objective, on the other hand, generates a perspective that is coherent to the context provided. For the query shown in Table 6, the baseline BART model incorrectly produces a supporting perspective to the

query, while the editorial or abstract presents the opposite stance. The model with the stance objective generates a perspective with a matching stance.

While we choose relevance and stance classification as the two auxiliary tasks in this case study, there exist many other candidate tasks that might be helpful in the setting. For instance, measuring the quality (Toledo et al., 2019), or more specifically persuasiveness (Carlile et al., 2018) of the perspective might be two viable options. As our study assumes that the abstract is provided for each editorial, the overall performance of perspective summarization will likely drop, if we use model-generated abstract instead of ground truth as input.

6 Related Work

6.1 Argumentation in News Editorials

News editorials have been studied as a resource for studying many argumentation-related tasks. Wilson and Wiebe (2003); Yu and Hatzivassiloglou (2003) use editorials for the study on sentiments and opinions. Later works (Reed et al., 2008; Bal and Saint-Dizier, 2009; Chow, 2016) shift focus on the argumentation structure within editorials, and their persuasiveness effect (Al Khatib et al., 2016; El Baff et al., 2020). A few other recent studies have explored argument quality (El Baff et al., 2018) and generation (El Baff et al., 2019) when using editorials as a resource.

Our proposed dataset and study focus on the interplay between elements of the argumentation structure presented in editorial articles. Unlike previous work, we study these elements as the abstractive instead of extractive summary from the news editorials.

6.2 Argument Generation

Most early efforts in argument generation, i.e. generating components in an argumentation structure, study rule-based synthesis methods based on argumentation theories (Reed et al., 1996; Zukerman et al., 2000). With the recent progress in neural, sequence to sequence text generation methods (Sutskever et al., 2014), a few studies have adapted such techniques for end-to-end argument generation. (Wang and Ling, 2016; Hua and Wang, 2018; Hua et al., 2019).

The task of perspective generation in this work closely relates to argument conclusion generation (Alshomary et al., 2020). Our study focuses on the setting where the target topic, or the *query*, is given as input to the generation model. Due to the implicit nature of the *perspectives* (Habernal et al., 2018), one key challenge to the task is keep the semantics of the perspective generated *truthful* to the abstract and editorial article. We approach this by measuring the compatibility of the perspective to the context along the dimensions of content salience (Bar-Haim et al., 2020) and stance correctness (Bar-Haim et al., 2017). Our multi-task generation approach conceptually resembles the work by Guo et al. (2018), where multiple auxiliary tasks is employed to improve the quality of the generated summary.

7 Conclusion

We present MULTIOPED an open-domain news editorial corpus that induces a number of argumentation-related tasks. The proposed dataset presents a few properties that are absent from existing datasets. First, the elements in the annotation structure are presented as abstraction over the text in editorial, as such elements usually exist implicitly in editorials. Second, as the pairs of editorials are aligned by topic, and exhibit opposing stance to each other, such structure allows for studies on cross-document argumentation structure. Third, the dataset allows for the study of multiple argumentation-related tasks together.

To demonstrate the power of having multiple related tasks in a single high-quality dataset, we study the problem of perspective summarization in a multi-task learning setting. Our analysis shows that modeling stance and relevance classification jointly with the summarization task improves the overall quality of the perspective generated.

In future work, we hope to utilize the corpus to improve the multi-task framework for perspective summarization. As we set aside the problem of abstract generation in our case study, we would also like to identify the challenges and potential solution to the problem. We hope that MULTIOPED presents opportunities and challenges to future research in argumentation.

Acknowledgments

The authors would like to thank Daniel Ravner, the CEO of www.theperspective.com, for kindly granting access to data from the site for academic research. This work was supported in part by a Focused Award from Google, and a gift from Tencent.

References

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A News Editorial Corpus for Mining Argumentation Strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443.

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.

Bal Krishna Bal and Patrick Saint-Dizier. 2009. Towards an analysis of argumentation structure and the strength of arguments in news editorials. The Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. 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.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Marisa Chow. 2016. Argument identification in Chinese editorials. In *Proceedings of the NAACL Student Research Workshop*, pages 16–21, San Diego, California. Association for Computational Linguistics.

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

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242, Lisbon, Portugal. Association for Computational Linguistics.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. 2020. Corpus wide argument mining-a working solution. In *AAAI*, pages 7683–7691.

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

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697, Melbourne, Australia. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan. Asian Federation of Natural Language Processing.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

Johannes Kiesel, Khalid Al-Khatib, Matthias Hagen, and Benno Stein. 2015. A shared task on argumentation mining in newspaper editorials. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 35–38, Denver, CO. Association for Computational Linguistics.

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

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

Chris Reed, Derek Long, and Maria Fox. 1996. An architecture for argumentative dialogue planning. In *International Conference on Formal and Applied Practical Reasoning*, pages 555–566. Springer.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *LREC 2008*.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Theresa Wilson and Janyce Wiebe. 2003. Annotating opinions in the world press. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, pages 13–22.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136.

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*.

Ingrid Zukerman, Richard McConachy, and Sarah George. 2000. Using argumentation strategies in automated argument generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 55–62, Mitzpe Ramon, Israel. Association for Computational Linguistics.

A Training Details

A.1 Experiment Settings

In this section, we describe our experiment settings in more details for reproducibility. We randomly split the MULTIOPED dataset into 70%/10%/20% splits for training, validation, and testing, respectively. We train each system for 6 epochs using the same training set, and use the validation set to find the best α_1 and α_2. We report the test set results in Table 3. All test set results are averaged results using three different random initilizations. The approximate training time for a system trained with 6 epochs on a 12GB GPU is less than an hour.

For results shown in Table 3, we use $\alpha_1 = 30$ in + REL, $\alpha_2 = 1$ in + Stance, and $\alpha_1 = 1, \alpha_2 = 1$ in + REL & STANCE, as they achieved the best results in the dev set during hyperperamater tuning A.2.

A.1.1 BART

BART pre-trained model has been proved effective on text generation, question answering, and summarization tasks (Lewis et al., 2020). Given the limited size of our dataset, we finetune on BART to transfer learn from the large amount of data it was pre-trained on. We use BART base with 6 encoder and decoder layers with hidden size of 768. We use AdamW with learning rate 3e-5 as our optimizer.

A.1.2 BERT Relevance Classifier

BERT pre-trained model has demonstrated its power in question answering, language inference, and text classification tasks (Devlin et al., 2019). We finetune BERT-mini on *PERSPECTRUM* dataset first on relevance classification task (Chen et al., 2019), and then finetune it on our dataset, yielding an accuracy of 92% on evaluation set (20% of the data). The finetuned BERT model has 4 layers and hidden size of 256.

A.1.3 BERT Stance Classifier

As the Relevance Classifier, we also finetune it on *PERSPECTRUM* dataset before training on our dataset. However, we use a larger BERT model with 8 layers and hidden size of 768 since it is a slightly more difficult task than the relevance task. For the 30 queries (2% of the whole corpus) that are labeled as open-ended questions and do not constitute a clear stance, we exclude them from the training of BERT stance classifier. Our classifier eventually achieves 75% accuracy in the 20% evaluation set.

A.2 Hyperparameter Tuning

To control the degree we penalize our model using auxiliary loss, we introduce hyperparameters α. We use the 10% validation set to choose our best parameters. We tune α for different values from 0.1 to 50 to examine its affect on our model. We choose the best α according to their relevance and stance scores, and if there is a tie, we select the one with the higher ROUGE2 score. Table 7, Table 8, and Table 9 show the validation set results for tuning the BART+Rel, BART+Stance, and BART+Rel & Stance systems, respectively.

α_1	Relevance Score
0.1	90.0
1	91.43
5	88.21
15	92.5
30	**92.5**
50	92.5

Table 7: Tuning α_1 for BART + Rel. Here we choose $\alpha_1 = 30$ since it has the highest ROUGE2 score.

α_2	Stance Score
0.1	72.14
1	**72.50**
5	64.64
15	64.29
30	60.36
50	62.50

Table 8: Tuning α_2 for BART + Stance

α_2	α_1	Relevance	Stance
0.1	1	92.86	71.07
0.1	5	88.21	63.93
0.1	50	92.5	64.64
1	**1**	**93.93**	68.93
1	5	90.71	**71.79**
1	50	93.57	70.36

Table 9: Tuning α_1 and α_2 for BART + Rel & Stance. Here we choose ($\alpha_1 = 1, \alpha_2 = 1$) over ($\alpha_1 = 5, \alpha_2 = 1$) since it has a higher ROUGE2 score.

A.3 Measure of Agreement

We use Cohen's and Fleiss kappa to measure the inter-rater agreement among annotators (Fleiss and Cohen, 1973). We calculate Cohen's kappa agree-

ment when there are only two raters, and Fleiss's kappa when there are more than two.

Cohen's Kappa: Let n be the number of instances to be labeled by A and B two raters. g is the number of distinct categories, and f_{ij} denotes the frequency of the number of subjects with the ith categorical response for rater A and the jth categorical response for rater Y. The kappa agreement is then calculated as

$$p_0 = \frac{1}{n}\sum_{i=1}^{g} f_{ii}$$
$$p_e = \frac{1}{n^2}\sum_{i=1}^{g} f_{i+}f_{+i}$$
$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

where f_{i+} is the total for the ith row f_{+i} and is the total for the ith column in the frequency table.

Fleiss Kappa: Let N be the total number of subjects, let n be the number of ratings per subject, and let k be the number of categories into which assignments are made. Let n_{ij} represent the number of raters who assigned the i-th subject to the j-th category. The kappa agreement is calculated as

$$p_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$$
$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} p_i$$
$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$$
$$P_e = \sum_{j=1}^{k} p_j^2$$
$$\kappa = \frac{\bar{P} - P_e}{1 - P_e}$$

B Example Screenshots from ThePerspective Website and MTurk Annotation Interface

In this section, we show example screenshots from the website where we extract the data, www.theperspective.com/perspectives, and example screenshots of our data annotation process.

For the three data annotation tasks using Mechanical Turk, stance annotation, implicit reference resolution, and topic annotation, we present the Turkers with definitions and instructions of the tasks that we require them to do, and 3-6 example questions with our expected answers. We ask them to read and comprehend our instructions before annotating, and use random control sets to filter out invalid annotations. More details can be found in the screenshots below.

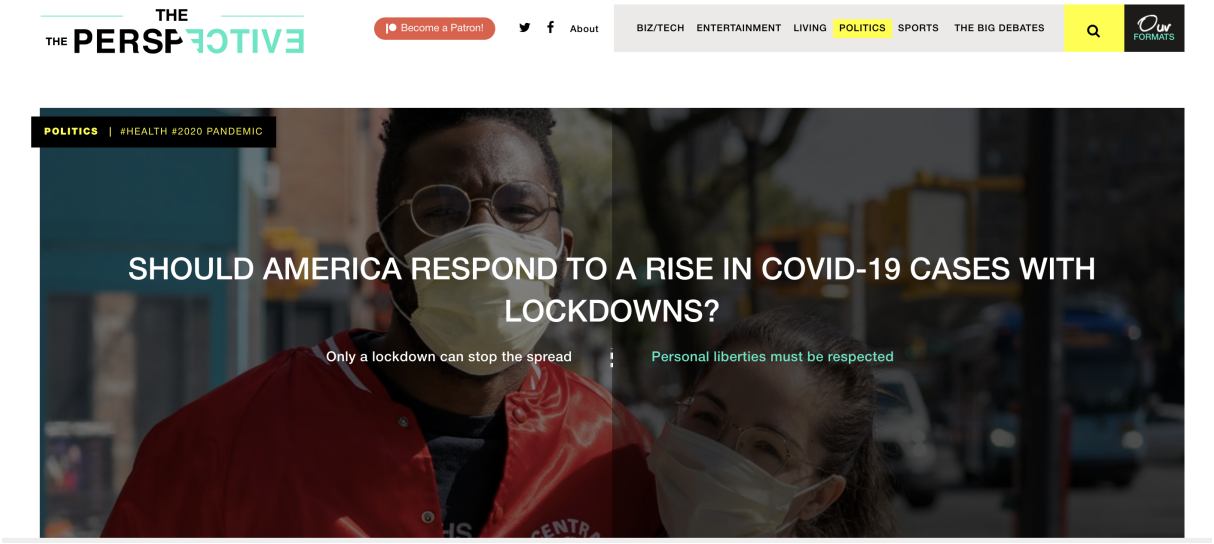

Figure 6: An example of a query and its two perspectives in ThePerspective website

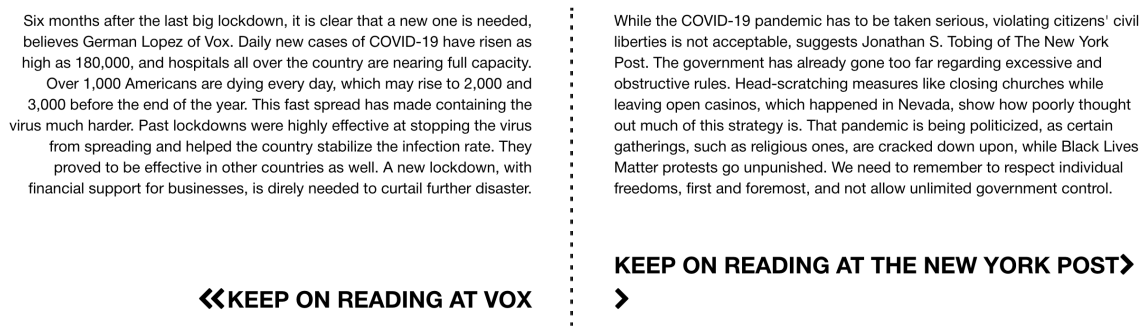

Figure 7: An exmaple of two abstracts and their links to news editorials in ThePerspective website

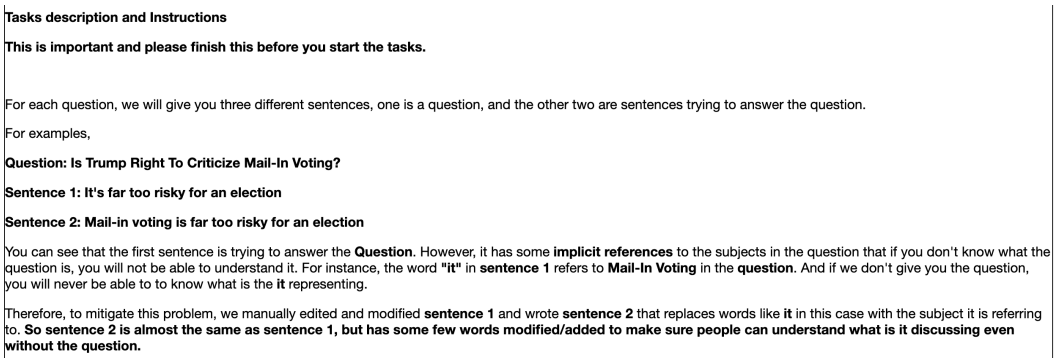

Figure 8: A screenshot of the MTurk annotation instruction for implicit reference resolution part1

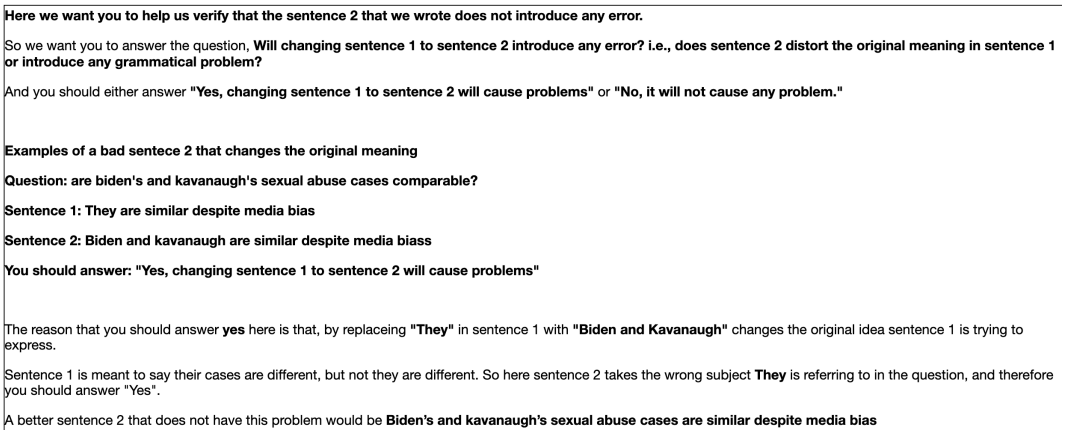

Figure 9: A screenshot of the MTurk annotation instruction for implicit reference resolution part2

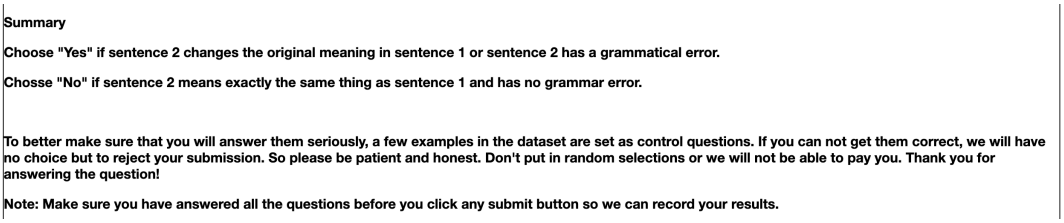

Figure 10: A screenshot of the MTurk annotation instruction for implicit reference resolution part3

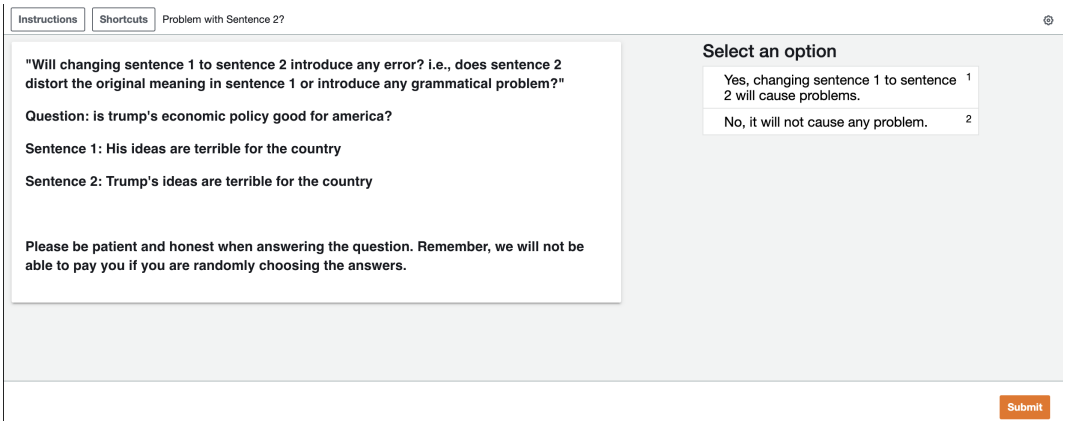

Figure 11: A screenshot of the MTurk annotation example for implicit reference resolution

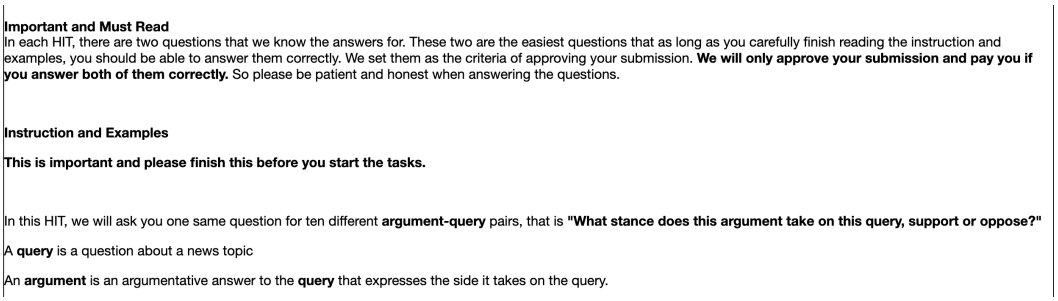

Figure 12: A screenshot of the MTurk annotation instruction for stance annotation part1

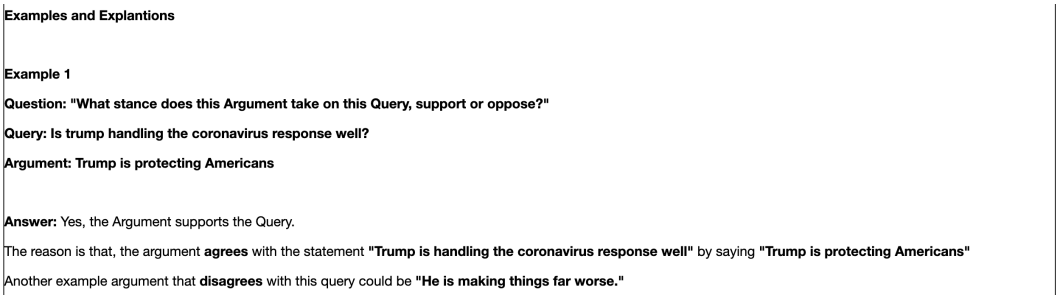

Figure 13: A screenshot of the MTurk annotation instruction for stance annotation part2

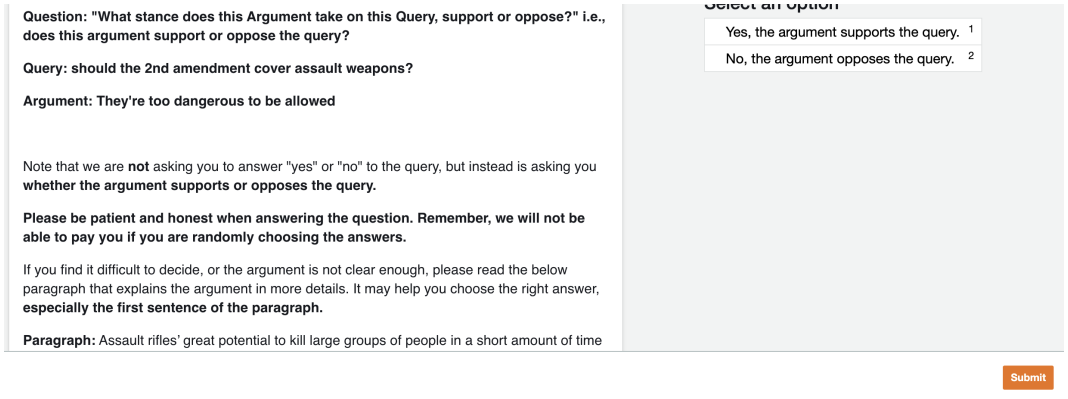

Figure 14: A screenshot of the MTurk annotation example for stance annotation. Note that there will be more sentences shown in the Paragraph line if the user scrolls down.

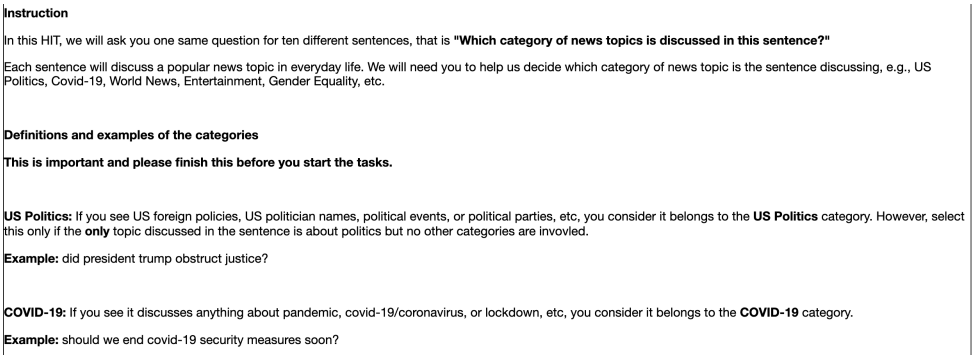

Figure 15: A screenshot of the MTurk annotation instruction for topic annotation part1

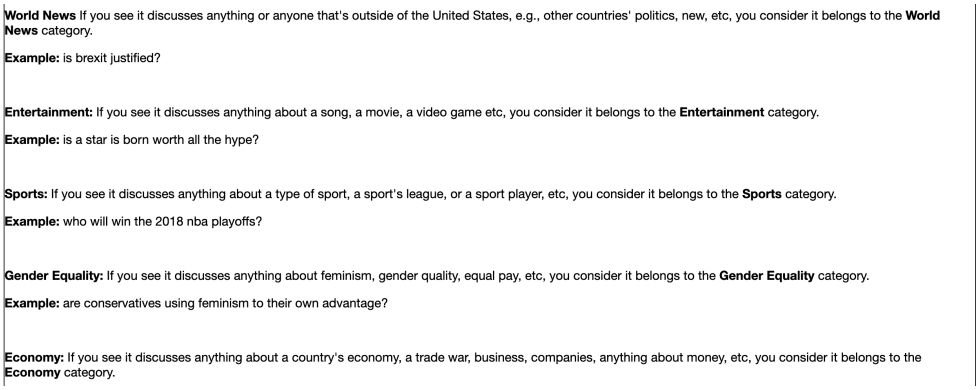

Figure 16: A screenshot of the MTurk annotation instruction for topic annotation part2

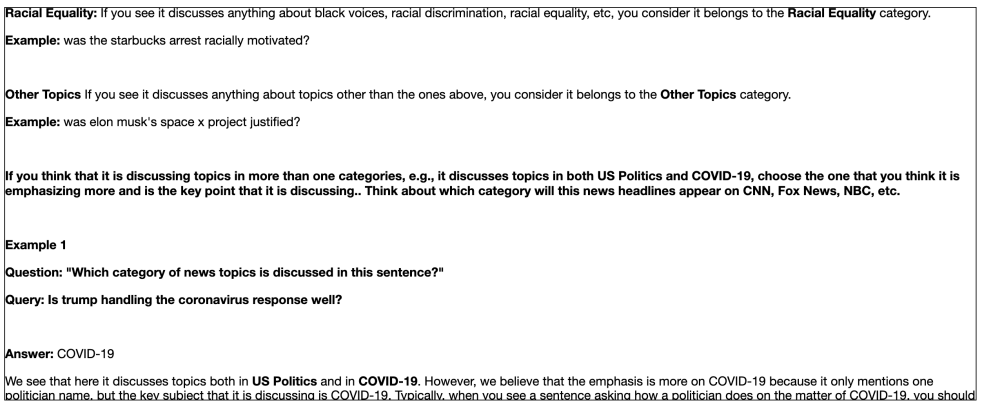

Figure 17: A screenshot of the MTurk annotation instruction for topic annotation part3

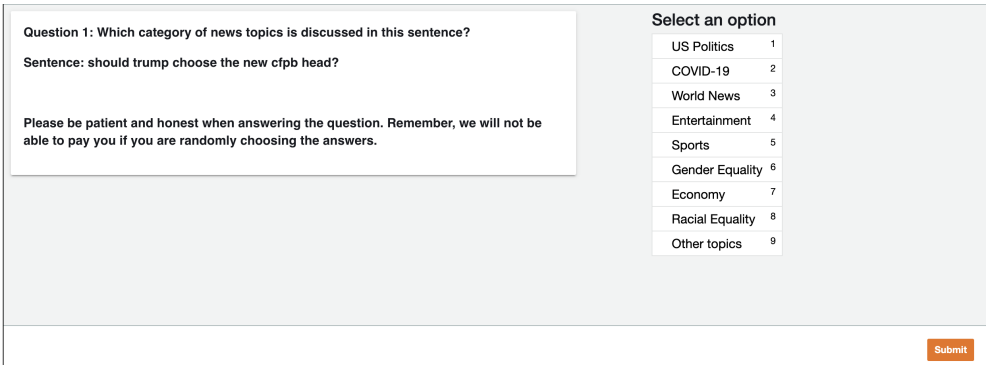

Figure 18: A screenshot of the MTurk annotation example for topic annotation