Out of the Echo Chamber:
Detecting Countering Debate Speeches

Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav,
Michal Jacovi, Ranit Aharonov and Noam Slonim
IBM Research
{matano,yonatanb,michal.jacovi,noams}@il.ibm.com
{assaf.toledo,dan.lahav, ranit.aharonov}@ibm.com

Abstract

An educated and informed consumption of media content has become a challenge in modern times. With the shift from traditional news outlets to social media and similar venues, a major concern is that readers are becoming encapsulated in "echo chambers" and may fall prey to fake news and disinformation, lacking easy access to dissenting views. We suggest a novel task aiming to alleviate some of these concerns – that of detecting articles that most effectively counter the arguments – and not just the stance – made in a given text. We study this problem in the context of debate speeches. Given such a speech, we aim to identify, from among a set of speeches on the same topic and with an opposing stance, the ones that directly counter it. We provide a large dataset of 3685 such speeches (in English), annotated for this relation, which hopefully would be of general interest to the NLP community. We explore several algorithms addressing this task, and while some are successful, all fall short of expert human performance, suggesting room for further research. All data collected during this work is freely available for research[1].

1 Introduction

Recently, a publication on Quantum Computing described a quantum computer swiftly performing a task that arguably would require 10,000 years to be solved by a classical computer (Arute et al., 2019). A non-expert reader is likely to consider this claim as a hard-proven fact, especially due to the credibility of the venue in which this publication appeared. Shortly afterwards, a contesting blog written by other experts in that field[2]

argued, among other things, that the aforementioned problem can be simulated on a classical computer, using proper optimizations, in 2.5 days. Clearly, out of potentially many texts questioning the promise of Quantum Computers (e.g. Kalai (2019)), making readers of the former publication aware of that specific blog post, which directly contests the claims argued in that publication, will provide them with a more informed view on the issue.

Broadly, argumentative texts, such as articles that support a certain viewpoint, often lack arguments contesting that viewpoint. This may be because those contesting arguments are not known to the author of the text, as they might not even have been raised at the time of writing. Alternatively, authors may also deliberately ignore certain known arguments, which might undermine their argumentative goal. Regardless of the reason, this issue places readers at a disadvantage. Lacking familiarity with opposing views that *specifically challenge a given perspective*, may lead to uninformed decisions or establishing opinions based on partial or biased information. Therefore, there is merit to developing a system that can automatically detect such opposing views.

Motivated by this scenario, we propose a novel natural language understanding task: Given an input text and a corpus, retrieve from that corpus a *counter text* which includes arguments contesting the arguments raised in the input text. While contemporary systems allow fetching texts on a given topic, and can employ existing tools to discern its stance – and so identify texts with an opposing view – they lack the nuance to identify the counter text which directly contests the arguments raised in the input text.

The potential use-cases of the proposed system exist in several domains. In politics, it can present counters to partisan texts, thus promot-

---

ing more informed and balanced views on existing controversies. In social media, it can alleviate the bias caused by the "echo chamber" phenomenon (Garimella et al., 2018), by introducing opposing views. And in the financial domain, it can potentially help analysts find relevant counter-texts to predictions and claims made in earning calls. It may also help authors to better present their stance, by challenging them with counter texts during their writing process. Lastly, it may aid researches to examine relevant citations by annotating which papers, out of potentially many, hold opposing views. Note, however, that this paper focuses on counter text detection - a useful tool for these worthy goals, but not a complete solution.

To pursue the aforementioned task, one needs a corresponding benchmark data, that would serve for training and evaluating the performance of an automatic system. For example, one may start with an opinion article, find a set of opinion articles on the same topic with an opposing stance, and aim to detect those that most effectively counter the arguments raised in the opinion article we started with. This path represents a formidable challenge; for example, reliable annotation of long texts is notoriously difficult to obtain (Lavee et al., 2019a), to name just one reason out of many.

To overcome this issue, here we focus on a unique debate setup, in which the goal of one expert debater is to generate a coherent speech that counters the arguments raised in another speech by a fellow debater. Specifically, as part of Project Debater[3], we collected more than 3,600 debate speeches, each around four minutes long, recorded by professional debaters, on a wide variety of controversial topics, posed as debate *motions* (e.g. *we should ban gambling*). With this paper, we make this data available to the community at large. Each motion has a set of supporting speeches, and another set of opposing speeches, typically recorded in response to one – and only one – of the supporting speeches. Correspondingly, our task is defined as follows. Given a motion, a supporting speech, and a set of candidate opposing speeches discussing the same motion, identify the opposing speeches recorded in response to the supporting speech.

We analyze human performance on this challenging task, over a sample of speeches, and further report systematic results of a wide range of contemporary NLP models. Our analysis suggests that expert humans clearly outperform the examined automatic methods, by employing a potentially non-trivial mix of heuristics.

In summary, our main contributions are as follows: (1) Introducing a novel NLU task, of identifying the long argumentative text that best refutes a long argumentative text given as input. (2) Suggesting to simulate the proposed general task in a well-framed debate setup, in which one should identify the response speech(es) that rebuts a given supporting speech. (3) Sharing a large collection of more than 3,600 recorded debate speeches, that allow to train and evaluate automatic methods in our debate-setup task. (4) Providing empirical results for a variety of contemporary NLP models in this task. (5) Establishing the performance of humans in this task, conveying that expert humans currently outperform automatic methods.

2 Related Work

Most similar to our work is the task of retrieving the best counter argument to a single given argument (Wachsmuth et al., 2018), also within the debate domain. However, in that setting counterarguments may discuss different motions, or have the same stance towards one motion. In our setting, identifying speeches discussing the same motion can be done using existing NLP methods, and being of opposing stances may be explored with various sentiment analysis techniques. Our focus is on identifying the response to a supporting speech within a set of *opposing* speeches, *all discussing the same motion*. Other than the different setup, our task also handles a more complex premise – speeches which are substantially longer than any single argumentative unit, and include multiple such units.

An alternative to our approach is breaking the problem into three stages: (1) identifying specific arguments made in each debate speech; (2) establishing counterargument relations between such arguments found in different speeches; (3) choosing the best response speech based on these argument-level relations. The first subproblem has been recently explored in Mirkin et al. (2018b); Lavee et al. (2019b); Orbach et al. (2019). The second is related to a major research area within computational argumentation

---

[3]https://www.research.ibm.com/artificial-intelligence/project-debater/

(see recent surveys by Cabrio and Villata (2018); Lawrence and Reed (2019)). Such research includes detecting attack relations between arguments (Cabrio and Villata, 2012; Rosenthal and McKeown, 2015; Peldszus and Stede, 2015b; Cocarascu and Toni, 2017; Wachsmuth et al., 2018), modeling them (Sridhar et al., 2015), depicting these relations (Walker et al., 2012; Peldszus and Stede, 2015a; Musi et al., 2017), generating counter-arguments (Hua and Wang, 2018; Hua et al., 2019), and establishing a theoretical framework for engagement (Toulmin, 2003; Govier, 1991; Dung, 1995; Damer, 2009; Walton, 2009).

A major drawback of the above approach is that it requires a considerable labeling effort – the annotation of arguments mentioned within speeches – which has been shown to be a challenge (Lavee et al., 2019a). Another is that the methods in the above studies which focus on establishing relations at the individual argument level may be limited when aiming to evaluate the perspective of long texts. Specifically, a response speech may contain multiple arguments that relate to the supporting speech in different ways. For instance, the speaker in such a speech may choose to concede an argument, while still maintaining an opposite view. Therefore simply mapping argument level relations may fall short when trying to generalize and assess full speeches. Our task complements the above endeavors by facilitating a framework that would allow extending their granularity from the argument level to a full-text level. Also, our main motivation is different – detecting whole long counter speeches, and not the exact counter arguments within the counter speech. The latter, perhaps more challenging goal, is out of scope for this work.

New neural models have recently driven performance improvements across many NLP tasks (Devlin et al., 2018; Radford et al., 2018), surpassing the level of non-expert humans in a diverse set of benchmark tasks (Wang et al., 2018; McCann et al., 2018). To facilitate the progress of further research Wang et al. (2019) introduced a benchmark aiming to pose a new series of rigorous tests of language understanding which are challenging for cutting-edge NLP technologies. Our work is consistent with the motivation behind these benchmarks, as it suggests a challenging new NLU task, accompanied by a corresponding dataset and benchmarks.

The rise of deliberate disinformation, such as fake news, highlights the erosion in the credibility of consumed content (Lazer et al., 2018), and situations where one is exposed only to opinions that agree with their own, as captured by the notion of echo chambers, are becoming more prevalent (Garimella et al., 2018; Duseja and Jhamtani, 2019). The task proposed in this work seems timely in this context.

3 Data

We now detail the process of collecting the speeches, the structure of the dataset, and how it is used for our task.

Dataset structure   Each speech in the dataset discusses a single motion and is either a *supporting speech* – in which a single speaker is arguing in favor of the discussed motion, or an *opposing speech* – in which the speaker is arguing against the motion, typically in response to a supporting speech for that motion. As described below, debaters recording an opposing speech typically listen to a given recorded supporting speech, and then design and record their own speech in response to it. This *counter speech* is either *explicit* – including a rebuttal part in which the speaker directly addresses arguments raised in the rebutted speech, or *implicit* – including no such dedicated rebuttal section, but tacitly relating to the issues raised in the supporting speech they respond to. The data contains multiple counter speeches to each supporting speech, among which some, none or all may be explicit or implicit. Figure 1 depicts the structure of this dataset. Examples of one explicit and one implicit counter speeches are included in the Appendix.

Recording speeches   The supporting speeches were produced by a team of professional debaters, using a procedure similar to the one described in Mirkin et al. (2018a): The debaters were each given a list of motions, accompanied by relevant background materials (taken from an online resource such as Wikipedia). They were allowed ten minutes of preparation time to review a motion's background material, after which they recorded a speech arguing in favor of that motion, which was around four minutes long. Through this process, 1797 supporting speeches were recorded, discussing 460 motions.
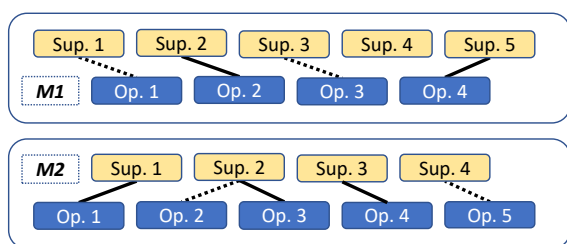
To record an opposing speech, the debaters were

Figure 1: The speeches data structure for two motions (*M1* and *M2*): Each motion has several supporting (**Sup.**) and opposing (**Op.**) speeches. Opposing speeches which constitute an **explicit/implicit** counter speech to a supporting speech are connected to it with a **solid/dashed** line. In the data, each supporting speech has zero or more counters, and each opposing speech is the counter of at most one supporting speech.

first given ten minutes to review the background material for the motion, as in the recording of a supporting speech. Then, they listened to a supporting speech (recorded by a fellow debater) and recorded a counter speech of similar length. Due to different debate styles popular in different parts of the world, some debaters recorded explicit counter speeches while others recorded implicit ones. To expedite the pace of the recording process, towards its end, few opposing speeches were recorded without requiring the debater to respond to a specific supporting speech. Instead, the debaters were instructed to think of supporting arguments themselves, and respond to these arguments. In total, 1887 opposing speeches were recorded: 348 are explicit counters, 1389 are implicit, and the other 150 are not the counter speech of any supporting speech. The full guidelines used by the debaters during the recordings are included in the Appendix.

The recorded audios were automatically transcribed into text using Watson's off-the-shelf Automatic Speech to Text (STT)[4]. Human transcribers listened to the recorded speeches, and manually corrected any errors found in the transcript texts produced by the STT system. On average, each speech transcript contains 28.2 sentences, and averages 738.6 tokens in length.

For the purpose of this work, the manually-corrected transcripts are used. The full data of 3685 speeches, including the recorded audios, the STT system outputs and the manually-corrected

transcripts are available on our website[5]. For comparison, the previous release of Project Debater's speeches dataset (Lavee et al., 2019b) included a smaller subset of 400 speeches. Further details on the format of the full data and the recordings process are available in Mirkin et al. (2018a).

Usage   As noted above, our task input is comprised from a supporting speech and several candidate opposing speeches all discussing the same motion. Some candidates are counters of the supporting speech, and others are typically counters of *other* supporting speeches for the same motion. The goal is to identify those counter speeches made in response to the supporting speech. Opposing speeches produced by the speaker of the supporting speech were excluded from the candidates set, as in the real world it is unexpected for one to simultaneously support both sides of a discussion.

4   Human Performance

Recently, with deep learning techniques achieving human performance on several NLU tasks, and even surpassing it, there is growing interest in raising the bar (Wang et al., 2019). That is, to facilitate advancing NLU beyond the current state-of-the-art, there is a need for novel tasks which are solvable by humans, yet challenging for automatic methods. To assess our proposed task in this context, we performed an annotation experiment, as described below.

Setup   Each question presented one supporting speech and between 3 to 5 candidate opposing speeches, all discussing the same motion. Annotators were instructed to read the speeches, and select one opposing speech which they thought was a counter speech of the supporting speech. When they could not identify such a counter, they were asked to guess and mention that they had done so.

60 questions were randomly sampled and given to 3 English-proficient expert annotators, who have successfully worked with our team in other past annotation experiments. Following their choice of a counter speech, they were asked to explain their choice in free form language.

Following this step, one of the authors read the explanations provided by the experts and formed

---

[4]https://www.ibm.com/cloud/
watson-speech-to-text

[5]https://www.research.ibm.com/
haifa/dept/vst/debating_data.shtml#
DebateSpeechAnalysis

	All		Explicit		Implicit	
	A	**R**	**A**	**R**	**A**	**R**
Ex	85.6	31.4	91.7	37.4	76.4	30.2
Cr	48.9	28.4	55.6	29.5	38.9	26.6

Table 1: Annotation results showing, for each ***annotation setting***, the average accuracy (**A**) obtained by the experts (**Ex**) and crowd annotators (**Cr**), along with the accuracy of randomly guessing the answer (**R**).
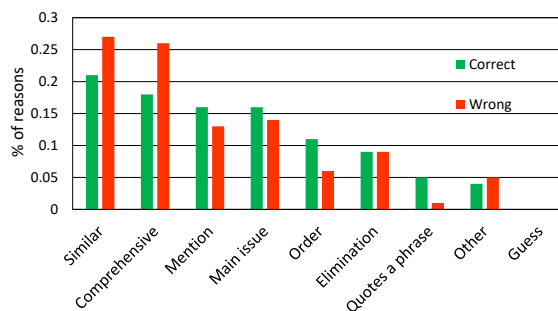

Figure 2: The distribution of reasons for the correct and wrong answers of crowd annotators (who overall had accuracy $\geq 60\%$).

a set of reason categories. Then, another 60 questions were sampled and given to 3 crowd annotators, using the *Figure-Eight*[6] crowdsourcing platform. The crowd annotators were from a dedicated group which regularly participates in annotations done by our team. After choosing a counter speech, they were instructed to choose the reason (or multiple reasons) for their choice from the set of reason categories. The crowd payment was set to 2.5$ per question. To encourage thorough work, a post-processing bonus was given for each correct answer, doubling that pay.

The full guidelines given to the expert and crowd annotators are provided in the Appendix.

Results   Performance was evaluated by calculating the accuracy of each annotator, and averaging over annotators. These results are presented in Table 1. Overall, the experts obtained an average accuracy of 86% (**Ex** row), considerably better than randomly guessing the answer which yielded an accuracy of 31%. The accuracy of the crowd annotators (**Cr**) was lower, yet distinctly better than random. This suggests that the task is difficult, and may require a level of dedication or expertise beyond what is common for crowd-annotators. Fortunately, the dataset is constructed in such a way that human annotation is not required to label it - it is clear by design which opposing speech counters which supporting speech.

To establish whether identifying explicit counters is easier than identifying implicit ones, the average annotator accuracy was separately computed for these two types. Noteworthy, the accuracy of the experts drops from a near perfect score of 92% on questions with an explicit true counter, to 76% on questions with an implicit one. Some of the drop may be explained by the smaller chance of guessing the correct answer at random over

this set, but not all[7]. This suggests that, as may be expected, identifying implicit counter speeches is more challenging than identifying an explicit counter. Still, the performance of both types of annotators, over both types of speeches, was better than random.

Reasons analysis   The explanations provided by the experts revealed several best-practices for this task, which we categorized as follows: The true counter speech **quote**s a phrase from the supporting speech; **mention**s a specific case or argument from the supporting speech; is more **comprehensive** and addresses more issues raised in the supporting speech than the other candidates; addresses those issues in the same **order** as they appear in the supporting speech; discusses **similar** issues; deals with the **main issue** raised in the supporting speech. Another reason was **elimination** – discarding the other candidates since they responded to issues or arguments which were *not* raised in the supporting speech. The last two categories were **guess** and **other** (which required writing a reason in free form language).

Focusing on crowd annotators who did the task relatively well (accuracy $\geq 60\%$), Figure 2 presents the distribution of the reasons they gave for their answers, separated between cases when they were correct and when they were wrong. Overall, the reasons distribution suggests that correctly solving this task requires balancing between the various heuristics. While some of these reasons, such as **similarity**, correspond to existing algorithmic ideas, others (e.g. **order** or **main issue**)

---

[7]Suppose that when answering, annotators answer correctly a fraction f of the time, and guess $1 - f$ of the time, with probability of success equal to the random baseline. Then in the explicit case $f = 0.87$ and in the implicit $f = 0.67$.

could inspire future research.

5 Counter Speech Identification

Having established that experts perform well on this task, the question remains whether present NLP methods can match that performance.

5.1 Setup

Data A supporting speech was included in the experiments if (a) there was an opposing speech addressing it; and (b) there was at least one additional opposing speech discussing its motion which was produced either in response to *another* supporting speech, or without responding to any specific supporting speech. Supporting speeches not meeting these criteria were excluded from the analysis. With these criteria, the data used in the experiments comprised 1102 supporting speeches and 1708 opposing speeches, pertaining to 329 motions.

Split The motions were randomly split into train (60%), validation (20%) and test (20%) sets, and their speeches were partitioned accordingly.

Settings To separately evaluate the ability to detect explicit and implicit counters, the experiments were performed in three settings. The first utilized the entire data – given a supporting speech, all of the opposing speeches discussing its motion were considered as candidate counters. In the second setting, the true counter speeches were limited to explicit counters. Supporting speeches without any explicit counter were excluded. Similarly, in the last setting, the true counter speeches were limited to implicit counters, and supporting speeches without such counters were excluded. For example, a supporting speech with one explicit counter, one implicit counter and whose motion is associated with two other opposing speeches (which are not its counters), is considered with all four opposing speech candidates in the first setting and three such candidates in the second and third settings - the two non-counters and the one counter of the type relevant to the setting. Table 2 details the statistics of each data split and experimental setting.

Evaluation The methods described next score each of the candidate counters. We report the average accuracy of the top predictions (**A**) and the average mean reciprocal rank (**M**), defined as $1/r$ where r is the highest rank of a true counter.

5.2 Methods

Document similarity Our first method represented speeches as bag-of-terms vectors, where terms are stemmed unigrams appearing in at least 1% of the speech-pairs in the training set, and the term counts are normalized by the total count of terms in the speech. Given two vectors, their similarity was computed using the Cosine similarity (*Cos*) or the inverse Jensen-Shannon divergence (*JS*).

Similarity and Dissimilarity Wachsmuth et al. (2018) presented a method for retrieving the best counter argument to a given argument, based on capturing the similarity and dissimilarity between an argument and its counter. At its core, their method is based on two similarity measures between pairs of texts: (i) A *word-based* similarity, which is defined by the inverse Manhattan distance between the normalized term frequency vectors of the texts (where terms were as mentioned above); (ii) An *embeddings-based* similarity which used pretrained ConceptNet Numberbatch word embeddings (Speer et al., 2017) to represent the words of the texts, averaged those embeddings to obtain a vector representing each text, and calculated the inverse Word Mover's distance (Kusner et al., 2015) between these vectors.

Previously, these measures were used to predict the relations between a pair of argumentative units. Since our speeches may contain multiple arguments, and their location within the text is unknown, we defined this method at the speech level by considering every supporting speech sentence and every candidate counter speech sentence. For each measure, the similarities of one supporting speech sentence to all candidate counter speech sentences were aggregated by applying a function *f*, yielding a sentence-to-speech similarity. These sentence-to-speech similarities were aggregated using another function *g*, yielding a speech-to-speech similarity. We denote these speech-to-speech measures by w_{fg} for word-based similarities and e_{fg} for embedding-based similarities. As aggregation functions, the maximum ($\uparrow$), minimum ($\downarrow$), average ($+$) and product ($\times$) were considered. For example, $w_{\uparrow+}$ denotes taking the maximal word-based similarity of each supporting speech sentence to all candidate counter speech sentences, and averaging those values.

Lastly, following Wachsmuth et al. (2018) once more, the similarity (*SD*) between a supporting

Setting	Train				Validation				Test			
	#Su	#Op	av. #C	Pos	#Su	#Op	av. #C	Pos	#Su	#Op	av. #C	Pos
All	649	1021	5.2	31%	218	340	5.2	31%	235	347	5.2	31%
Explicit	159	542	5.6	24%	58	208	6.0	22%	51	188	5.4	25%
Implicit	556	999	5.0	30%	194	337	5.0	30%	201	343	5.1	30%

Table 2: Data split statistics for each ***experimental setting*** (see §5.1): The number of supporting (**#Su**) and opposing (**#Op**) speeches, the average number of candidate counter speeches for each supporting speech (**av. #C**), and the percentage of those candidates which are counter speeches (**Pos**).

speech and a candidate counter is defined as

$$\alpha \cdot sim - (1 - \alpha) \cdot dissim$$

where *sim* and *dissim* are of the form $w_{fg} + e_{fg}$, both *f* and *g* are aggregation functions, $sim \neq dissim$ and α is a weighting factor. In this scoring model *sim* aims to capture topic similarity, whereas subtracting *dissim* seeks to capture the dissimilarity between arguments from opposing stances. Admittedly, this method is more appropriate for some of the settings explored in Wachsmuth et al. (2018), in which the candidate counter arguments to a given argument may be discussing other topics, and their stance towards the discussed topic is unknown. We include their method here for completeness, and to allow a comparison to their work.

The hyper-parameters, namely, the aggregation functions and the value of α (from the range $\{1, 0.9, 0.8\}$ used by Wachsmuth et al. (2018)) were tuned on the validation set. An additional variant (*SD-e*) based solely on the embeddings-based similarity was also considered, since it carries the advantage of not requiring any vocabulary to be derived from the training set. This allowed tuning the hyper-parameters on a larger set comprised from both the training and validation sets.

BERT Devlin et al. (2018) presented the BERT framework which was pre-trained on the masked language model and next sentence prediction tasks. Assuming that an argument and its counter are coherent as consecutive sentences, and that the first sentences of the candidate speech reference the last sentences of the supporting speech, those parts were scored using the pre-trained next-sentence prediction model with (*BERT-T*) and without (*BERT*) fine-tuning. The considered sentences from each speech were limited to at most 100 words, since the pre-trained model is limited to 512 word pieces (assuming about two word

pieces per word). Specifically, from the first speech we took the greatest number of sentences from the end of the speech such that their total length was less than 100 words, and similarly for the second speech for its starting sentences. For fine-tuning, we used the supporting speeches with each of their true counter speeches as positive sentence pairs, and added an equal number of negative pairs where the supporting speech appears with a randomly sampled opposing speech that is not its counter.

ngram-based The methods described so far assign a score to a supporting speech and a candidate counter without considering the other candidates. Using that content can aid in detecting key phrases or arguments which best characterize the connection between the supporting speech and its counter – these are the ones which are shared between those speeches and are not mentioned in any of the other candidates. Having many such phrases or arguments may be an indication that a candidate is a true counter speech. Indeed, the **quote** and **mention** reason categories account for more than 20% of the reasons selected by the crowd annotators when answering correctly (see Table 2).

To capture this intuition, ngrams containing between 2 to 4 tokens were extracted from each speech. Those containing stopwords, and those fully contained within longer ngrams, were removed. The set of ngrams which appear in both the supporting speech and the candidate – but not in any of the other candidates – was calculated, and the total length of the ngrams it contains was used as the score of the candidate (*ngrs*).

Mutual Information The speeches were represented as bag-of-terms binary vectors, where the terms are stemmed unigrams (excluding stopwords). Each candidate counter was scored using the mutual information between its vector and the

vector of the supporting speech (*MI*).

In addition, the mutual information between those vectors, conditioned by the presence of terms in the other candidate counters (*c-MI*), was calculated as follows. Let v_s be a vector representing a supporting speech and $\{v_c\}_{c=1}^n$ be a set of n vectors representing its candidate counters. Let c be such a candidate counter, and o_c represent the concatenation of the vectors of the *other* candidates excluding c. Let $v_{c|k}$ denote the vector of values from v_c at the indices where the entries of o_c are k (for $k = 1$ or 0), and let $v_{s|k}$ be defined similarly. Then, the conditional mutual information of the candidate c is given by

$$\sum_{k=0}^{1} p\left(k\right) I(v_{s|k}; v_{c|k})$$

where $p\left(k\right)$ is the percentage of entries of o_c with the value k, and $I(\cdot, \cdot)$ is mutual information. Intuitively, this measure aims to quantify the information shared between a supporting speech and a candidate, after observing the content of all other candidates, and thus is similar in spirit to the ngram-based method mentioned above.

5.3 Results

Table 3 presents the results obtained by the different methods in our three experimental settings. These results show that there is a large performance gap between the implicit and explicit settings – in favor of the latter – for all methods (except *BERT*), suggesting it is an easier setting. This is consistent with the results of our annotation experiment.

While the best performing methods (*JS* and *c-MI*) surpass the performance of individual crowd annotators (see Table 1), which testifies to the difficulty of the annotation task, the human experts clearly do better, suggesting there is still much room for improvement.

Error analysis We have manually analyzed the top 3 implicit and explicit speeches for which the differences in mutual information between the predicted counter speech and the true counter speech were the greatest. Analysis revealed that such counter speeches are characterized by argumentative material that is thematically similar to the material of the input speech. Depending on the use case, such results are not necessarily errors, since if the goal is to find relevant opposing content it is beneficial to present such speeches, even

Method	All		Explicit		Implicit	
	A	**M**	**A**	**M**	**A**	**M**
JS	**51.1**	**0.69**	**80.4**	**0.88**	41.2	0.62
c-MI	50.6	**0.69**	72.5	0.84	**42.7**	**0.63**
MI	48.5	0.68	68.6	0.81	40.7	0.62
ngrs	45.1	0.65	60.1	0.73	38.8	0.61
SD-e	42.1	0.63	60.8	0.76	35.3	0.58
Cos	40.0	0.62	49.0	0.70	35.8	0.58
BERT	36.4	0.57	21.6	0.44	33.7	0.57
SD	31.9	0.57	52.9	0.70	31.3	0.56
BERT-T	32.2	0.56	49.0	0.70	35.2	0.58
Rand	31%	–	25%	–	30%	–

Table 3: Experimental results on the test set for each **method** and *experimental setting*: The average accuracy of the top prediction (**A**) and the average mean reciprocal rank (**M**) of the true counter with the highest score. The methods are ordered by their **M** score in the *All* setting. The last row (**Rand**) shows the accuracy of the random baseline.

if they were not authored in response to the input speech. However, in some instances a thematically similar argument may be an irrelevant counter as arguments can share a theme without being opposing. For example, an input text may discuss an argument pertaining to the *rights* of a disenfranchised group, while the counter may revolve around *pragmatic outcomes* to the same disenfranchised group. While these arguments are likely to share the theme of disenfranchisement they are not necessarily opposing.

6 Further Research Potential

The data presented here was collected to facilitate the development of Project Debater, and we chose the novel counter speech detection task to showcase this data and make it available to the community. However, the unique properties of our data – recorded speech which is more organized and carefully construed than everyday speech – make it interesting to revisit well-known NLP and NLU tasks. Several examples are listed below.

Author attribution: All speeches in the dataset are annotated for the debater who recorded them. It could be particularly interesting to study author attribution on our dataset as it contains persuasive language, relevant to opinion writing and social media. Additionally, we provide voice recordings

and transcripts for all speeches, enabling to study multi-modal methods for this task.

Topic identification: This is a well established research area which can be examined here in various aspects, including clustering speeches by topic, matching speeches to topics or extracting the topic of a speech without prior knowledge.

Whereas previous work often requires annotating the topics of texts and deducing a consensual label, in our data the topic of a speech is given by design.

Sentence ordering or local coherence: The sentence ordering task (Barzilay and Lapata, 2005) is concerned with organizing text in a coherent way and is especially relevant for natural language generation. Our dataset allows to study this using spoken natural language of a persuasive nature, that often relies on a careful development of an argumentative intent. The data also provides a unique opportunity to study the interplay between a coherent arrangement of language and the associated prosodic cues.

Other tasks The large scale of the dataset, over 200 hours of spoken content and their manually-corrected transcripts, enables its use in other speech-processing tasks that require such data. Some examples include speech-to-text, text-to-speech, and direct learning from speech of word (Chung and Glass, 2018) or sentence (Haque et al., 2019) embeddings. Such tasks often use large scale datasets of *read* content (e.g. Panayotov et al. (2015)), and our data allows their exploration in the context of spoken *spontenous* speech.

In addition, with further annotations of the dataset, it lends itself to other potential tasks. One example is the extraction of the main points of a speech or article. This can facilitate various downstream tasks, such as single document summarization in the context of spoken language. Another example is the annotation of named entities within the transcript texts, facilitating direct identification of those entities in the audio, similarly to the work of Ghannay et al. (2018).

7 Conclusions

We presented a novel NLU task of identifying a counter speech, which best counters an input speech, within a set of candidate counter speeches.

As previous studies have shown, and consistent with our own findings, obtaining data for such a task is difficult, especially considering that labeling at scale of full speeches is an arduous effort. To facilitate research of this problem, we recast the proposed general task in a defined debate setup and construct a corresponding benchmark data. We collected, and release as part of this work, more than 3,600 debate speeches annotated for the proposed task.

We presented baselines for the task, considering a variety of contemporary NLP models. The experiments suggest that the best results are achieved using Jensen–Shannon similarity, for speeches that contain explicit responses (accuracy of 80%) and using conditional mutual-information on speeches that respond to the input speech in an implicit way (accuracy of 43%).

We established the performance of humans on this task, showing that expert humans currently outperform automatic methods by a significant margin — attaining an accuracy of 92% on speeches with an explicit true counter, and 76% on speeches with an implicit one. Noteworthy is that some of the automatic methods outperform the results achieved by the crowd, suggesting that the task is difficult, and may require a level of expertise beyond layman-level.

The reported gap between the performance of expert humans and the results achieved by NLP models demonstrate room for further research. Future research may focus on the motivation we described, but may also utilize the large speeches corpus we release as part of this work to a variety of additional different endeavors.

Acknowledgments

References

Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. 2019. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*,

pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *IJCAI*, pages 5427–5433.

Yu-An Chung and James Glass. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark. Association for Computational Linguistics.

T. Edward Damer. 2009. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments*. Wadsworth/Cengage Laerning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Nikita Duseja and Harsh Jhamtani. 2019. A sociolinguistic study of online echo chambers on twitter. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 78–83, Minneapolis, Minnesota. Association for Computational Linguistics.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference (WWW)*, pages 913–922, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699. IEEE.

Trudy Govier. 1991. *A Practical Study of Argument*. Wadsworth Pub. Co.

Albert Haque, Michelle Guo, Prateek Verma, and Li Fei-Fei. 2019. Audio-linguistic embeddings for spoken sentences. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7355–7359. IEEE.

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.

Gil Kalai. 2019. The argument against quantum computers. *arXiv preprint arXiv:1908.02499*.

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*, pages 957–966. JMLR.org.

Tamar Lavee, Lili Kotlerman, Matan Orbach, Yonatan Bilu, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019a. Crowd-sourcing annotation of complex NLU tasks: A case study of argumentative content annotation. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 29–38, Hong Kong. Association for Computational Linguistics.

Tamar Lavee, Matan Orbach, Lili Kotlerman, Yoav Kantor, Shai Gretz, Lena Dankin, Shachar Mirkin, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2019b. Towards effective rebuttal: Listening comprehension using corpus-wide claim mining. *6th Workshop on Argument Mining*.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 0(ja):1–55.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. Cite arxiv:1806.08730.

Shachar Mirkin, Michal Jacovi, Tamar Lavee, Hong-Kwang Kuo, Samuel Thomas, Leslie Sager, Lili Kotlerman, Elad Venezian, and Noam Slonim. 2018a. A recorded debating dataset. In *Proceedings of LREC*.

Shachar Mirkin, Guy Moshkowich, Matan Orbach, Lili Kotlerman, Yoav Kantor, Tamar Lavee, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2018b. Listening comprehension over argumentative content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 719–724. Association for Computational Linguistics.

Elena Musi, Debanjan Ghosh, Mark Aakhus, Smaranda Muresan, and Nina Wacholder. 2017. Building an ontology of (dis) agreement space for argument mining. *SIGDIAL/SEMDIAL 2017 Joint Session on Negotiation Dialog (position paper)*.

Matan Orbach, Yonatan Bilu, Ariel Gera, Yoav Kantor, Lena Dankin, Tamar Lavee, Lili Kotlerman, Shachar Mirkin, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. A dataset of general-purpose rebuttal. 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 5590–5600, Hong Kong, China. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. pages 5206–5210.

Andreas Peldszus and Manfred Stede. 2015a. An annotated corpus of argumentative microtexts. In *Proceedings of the First Conference on Argumentation, Lisbon, Portugal*.

Andreas Peldszus and Manfred Stede. 2015b. Towards detecting counter-considerations in text. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 104–109.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners. *OpenAI Blog*.

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 116–125.

Stephen E. Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.

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

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817. Istanbul.

Douglas Walton. 2009. Objections, rebuttals and refutations. *Argument Cultures: Proceedings of the 8th OSSA Conference*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

A   Introduction

This appendix contains the guidelines used in all the data generation and annotation tasks described in the paper: 1) speech authorship guidelines, 2) identifying the response speech from a list of candidates, 3) identifying the response speech speech from a list of candidates and providing a reason.

Following the guidelines are two examples of full response speeches - an explicit counter speech and an implicit counter speech (see §3).

B   Speech Authoring Guidelines

For supporting speeches:

- Read the motion text and background.

- Prepare for 10 minutes while avoiding the use of external sources.

- Record a 4 min opening speech in a normal speaking pace.

For opposing speeches:

- Read the motion text and background.

- Prepare for 10 minutes while avoiding the use of external sources.

- Listen to the supporting speech.

- Immediately record a 4 min opening speech in a normal speaking pace.

- When recording your speech, please make sure to relate to the arguments raised in the government's opening speech; i.e., engage with them like you would have done in British Parliamentary debate style, or in any other kind of academic debate format.

C   Identify The Opposing Speech Guidelines

In this task you are given a motion and speech arguing in favor of that motion. It is then followed by 3-5 opposing speeches. One of those speeches was recorded in response to the first supporting speech. Please select the opposing speech which you think was recorded in response to the supporting speech. In addition, please write in your own language the reason for your choice.

Note that you MUST select exactly one opposing speech. If you aren't sure, take a guess, and

specify you had done so when detailing the reason for your choice. Some additional examples of valid reasons are "Both X and Y seemed reasonable choices, and X seemed more appropriate", "The supporting speech is talking about Z, as does the opposing speech", etc. No specific format is required for detailing the reason, but please do your best to be clear.

D   Identify The Opposing Speech (With Reasons) Guidelines

Overview

In this task you are given a controversial topic and a supporting speech arguing in favor of that topic. The supporting speech is followed by 3-5 opposing speeches. One of those opposing speeches was recorded in response to the supporting speech.

1. Select the opposing speech that was recorded in response to the supporting speech.

2. Select the reasons for your choice from a predefined list of reasons. You can select more than one reason.

3. Explain your choice, in your own words, in case the reason for your choice does not appear in the list.

Note that you MUST select exactly one opposing speech. If you aren't sure, take a guess, and specify you had done so when selecting the reason for your choice from the predefined list. When explaining your choice in your own words, no specific format is required – but please do your best to be clear.

Important Note

This task does not contain test questions, but your answers will be reviewed after the job is complete. We trust you to perform the task thoroughly, while carefully following the guidelines.

E   Example Speeches

Explicit counter speech: Opposing subsidies for higher education

"Before we begin there is something that, at least to me, was remained unclear in the mechanism, and that is the question of what exactly is going to get subsidized and what isn't. Do liberal arts studies or humanities studies are they going to get the

same full funding like computer science or engineering? We think that this is important because no matter what the answer is going to be, this raises some serious questions and difficulties but anyway, we're going to put that aside for now in the hope that government will make this clear in the next speech. So, side government is asking to convince us in the following things: a, education, no matter what age, is a basic right. B, if there is a basic right, then this automatically means that the government is also responsible to fully fund this. C, subsidizing, like a full subsidy of higher education, is going to be a smart investment that pays off in the long run, both economically and socially. We disagree with literally every one of those stages. Let's explain why. Firstly, on education in every age being a basic right. So government basically start by saying: look, we can all agree that primary education is a basic right and therefore, we must agree that that higher education is also a basic right. Now that is a logical leap. There are plenty of protections and special rights that we provide children but not adults. Children are protected, for instance, from criminal liability. And according to government's logic, if that is true, then this should also apply to adults. This is of course absurd. Specifically, the line that we cross between primary education to higher education isn't at all random. Primary education is a crucial condition to succeed in life, no matter what field you're going to to find yourself in. And that's what makes it a basic right. It is also a tool of the state to create a shared basis of knowledge to all of the citizens, sort of a way to shape the shared narrative and the collective identity of the nation. Higher education, on the other hand, isn't a crucial condition in plenty in like a lot of fields and and frankly, in the previous years, it is becoming less and less critical for success. In addition, there is also no element of like a a shared foundation here because everybody studies different things entirely, so no, this is not a basic right. Secondly, even if we were to agree that this is a basic right, this doesn't automatically mean that the government need to completely fully fund it. Food is also a basic right, right? And still the state helps you very partially and does not provide food for everyone free. We need to say this very clearly. The state already participates today in the funding of higher education in public institutes but in a partial way. We think that demanding that

it will provide for all of it is simply a misguided way of perceiving what the state's role is. Why isn't it enough to fund scholarships for less well-off students and continue collecting money from students that have no problem to fund themselves, for instance? And lastly, we get to the question of whether this is a smart investment. Now, as I have already hinted, higher education might have been critical for success in the market ten years ago or fifteen years ago, but the market is rapidly changing today and more and more of the most desired job places, for instance, in google or facebook, don't even demand a an academic title. We think that before we run off to spend billions of dollars on higher education free for everybody, then it's worth at least heavily considering these institutional changes, and that is something that side government isn't even considering. For all these reasons, please oppose."

Implicit counter speech: Opposing disbanding ASEAN

"We should not disband ASEAN. So, ASEAN is the association of southeast asian states. As the last speaker pointed out to you, it's made up of a group of states in southeast asia who are working together towards common goals of development. Three reasons why we should not disband it. First is about anti-colonialism. Recognize that for developing countries like the ones in ASEAN like malaysia, like indonesia, they have a few alternatives for who they can turn to as trade partners. You have major international trading countries like the states, like china, like EU countries, which historically have treated these countries in a very colonialist way. Most of the countries in ASEAN except for thailand were once colonized by european countries or the united states, and if you look back before that, they had a semi colonial relationship with china in many instances, such the relationship between vietnam and china. So we see that there's a history of abuse and mistreatment between these larger countries around the world the more powerful countries, and the ASEAN countries. We think that by working together, the ASEAN countries can ensure that they are a large enough economic bloc to prevent these major international powers who have historically come in and pushed them around, from dominating the region, in other words, ASEAN makes all of these countries that together are strong and

able to resist imperialist aggression or trade policy, and would all individually not be that powerful. It allows them to work together towards a common goal of independence and it reassures the independence of every member state from international oppression and dependence on one country for trade. Our second argument, is about why we think that fundamentally ASEAN increases development and that's the highest good in this round. So first, why is development the most important good? If you think about the quality of life in ASEAN countries, obviously it varies. People in malaysia for instance have like a middle income quality of life, people in vietnam are much poorer, but we think that overall everyone in all of these countries could still benefit from more development. We think that there is a moral imperative for states to seek out development for their citizens. Why is this so? So when we say a moral imperative we mean that states always have an obligation to seek this out. We think that because, any person would always choose to live in the most developed country possible so that they have the highest quality of life, those with the ability to do so, those who reap the benefits of developed life, because they're elites, should try to provide it to everyone else. To sort of do unto others as you would have them do unto you type of thinking. We see that, development is more likely with ASEAN. One, because countries have more access to trade partners and trade goods, so it's more like that they're able to specialize and develop industries that can then take advantage of other markets within ASEAN, and two, because of the access to economic development expertise. Recognize that many countries in ASEAN, are at different levels of development. Malaysia is pretty far along, some other countries are not as far along. So we tell you that people in ASEAN countries can study in other countries and learn about development and industry, and how other countries have been successful in the past, and use this in order to help their own home country. So at the end of the day, we help the people who are worse off in the world, some of them, some of these very poor people who live in ASEAN countries because we better access development so we shouldn't disband ASEAN. Our last argument is about peace in the region. Recognize that there are many potential sources of conflict within the southeast asia region. Some countries are more closely aligned with china so they see an advantage in china becoming more hegemonic, some countries are more aligned with the united states. Some countries are communists, some countries are capitalist. There's been conflict in the past over east timor, and there are other ethnic tensions boiling beneath the surface in many southeast asian countries. But one of the surest ways to prevent international conflict, is to tie everyone's interests together through trade. If everyone stands to get richer through peace and poorer through conflict, then it's much less likely that a war will break out in the region. So for that reason we think ASEAN is a tremendous tool for peace in southeast asia in the future. So because it's an anti colonial institution, because it promotes development, and because it will lead to peace in the region, we should not disband ASEAN thank you."