A Dataset of General-Purpose Rebuttal

Matan Orbach[1], Yonatan Bilu[1], Ariel Gera[1], Yoav Kantor[1],
Lena Dankin[1], Tamar Lavee[1], Lili Kotlerman[*2], Shachar Mirkin[*3],
Michal Jacovi[1], Ranit Aharonov[1] and Noam Slonim[1]
[1]IBM Research, [2]Intuition Robotics, [3]Digimind
{matano,yonatanb,arielge,yoavka,lenad,tamarl, michal.jacovi, ranita,noams}@il.ibm.com,
lili@intuitionrobotics.com, shachar.mirkin@digimind.com

Abstract

In Natural Language Understanding, the task of response generation is usually focused on responses to short texts, such as tweets or a turn in a dialog. Here we present a novel task of producing a critical response to a long argumentative text, and suggest a method based on general rebuttal arguments to address it. We do this in the context of the recently-suggested task of listening comprehension over argumentative content: given a speech on some specified topic, and a list of relevant arguments, the goal is to determine which of the arguments appear in the speech. The general rebuttals we describe here (written in English) overcome the need for topic-specific arguments to be provided, by proving to be applicable for a large set of topics. This allows creating responses beyond the scope of topics for which specific arguments are available. All data collected during this work is freely available for research[1].

1 Introduction

A key element in argumentation is rebuttal: the ability to contest an argument by presenting a counter-argument. It is an important skill, not easily learned, and valued in many fields such as politics and science. It is useful for advancing your own views and beliefs over opposing ones, but perhaps more importantly, it facilitates a critical examination of the views and beliefs that you hold. An automatic rebuttal system could therefore be useful whenever critical analysis of written or spoken content is required – be it an elementary school student writing an essay or a seasoned journalist composing an op-ed.

In the context of Natural Language Understanding, the study of rebuttal and counter-arguments

has focused on elucidating such relations between given arguments. Indeed, such "attack" relations are the foundation of *Argumentation Frameworks* (Dung, 1995); such frameworks have been one of the main objects of study in computational argumentation.

A related task, that of generating a response which need not be a rebuttal or even argumentative, has been the subject of much research, especially in the context of dialog systems, chat bots, and question answering. In this line of work the response typically follows a short input text, often only a sentence or two.

Here we suggest the task of producing a rebuttal in response to a long argumentative text. Specifically, we consider spoken speeches around four minutes long. In addition to being longer, and perhaps because they are so, these kinds of texts tend to include very general claims, often implicit in the text. As such, these claims may appear in varied contexts, and it may be feasible to compile a list of such claims independently of the speeches' topics.

For example, a concern that often comes up in debates about policy is that implementing the policy (or failing to do so) disproportionately harms minorities. This claim can be made to oppose school vouchers, to oppose voter registration laws or to support criminal justice reform. Moreover, for some such general claims, it is feasible to phrase a single rebuttal response which can fit many of the contexts in which the claim might be made. In the above example, such a response may talk about separating between the policy at hand, which should be adopted based on its merits, and the need to right historical wrongs, which should be pursued independently.

We envision an automatic rebuttal system based on this observation, which includes a manually curated General-Purpose Rebuttal Knowledge Base (GPR-KB) comprised of claims and matching re-

---

buttal responses. Given an argumentative text, the system would identify which claims from the GPR-KB are made in the text (explicitly or implicitly), and produce a rebuttal using the available counter-arguments. Clearly, many of the claims made in the text would not appear in the GPR-KB. The objective is therefore not to identify and rebut all arguments, but rather to identify and rebut *some* arguments, and construct a GPR-KB that facilitates that.

Such a system (based on the more elaborate CoPA modeling of Bilu et al., 2019) was indeed implemented as a key element in IBM's Project Debater rebuttal mechanism, and demonstrated during the live debate held between it and debating champion Harish Natarajan[2]. However, a rebuttal system of this nature may be of interest beyond the realm of debating technologies. For example, such a system may be instrumental in making media consumption a more critical process, by automatically challenging the consumer with counter-arguments. Similarly, it can be applicable in the education domain, stimulating critical thinking by prompting students with counter-arguments in response to (or during) essay-composition tasks.

The formation of a GPR-KB that is applicable to the real world poses several challenges. First, phrased claims must be both relevant to a variety of topics, and commonly used. Second, pre-written rebuttals should be effective and persuasive, even though they are created without prior knowledge of the context in which they are to be of use. We address these issues by turning to a domain in which a similar problem is solved by humans: the world of competitive debates. In these contests, successful debaters need to combine specific knowledge about the topic at hand, with general arguments that arise from the underlying principles of the debate. Their ability to use such general arguments for different topics lays the basis for using a GPR-KB as the one described above.

Accordingly, we asked an expert debater to create the initial GPR-KB by suggesting common claims and preparing matching rebuttals. The full process is detailed in §3.2.

To assess the usefulness of the suggested claims and rebuttals in the real world, we performed several steps of labeling on the dataset we constructed in Mirkin et al. (2018), containing spoken argu-

mentative content discussing controversial topics. Details of this process, along with an analysis showing the high coverage obtained by our knowledge base, are described in §4.

Another major challenge is the development of automatic methods for identifying whether knowledge-base claims are mentioned by speakers. We break this problem into a three-stage funnel – identifying whether: (i) a claim is relevant to the topic; (ii) the claim's stance aligns with that of the speaker; (iii) the claim was made by the speaker. We provide simple baseline results for this third step (§5). Interestingly, we observe that simply selecting the claim with the highest acceptance rate in the training data (without looking at the text) provides a challenging baseline.

The main contributions of this work are (i) the introduction of a novel task in NLU: producing rebuttal in response to a long argumentative text (ii) a manually constructed GPR-KB shared across multiple topics (iii) an additional layer of labeling to our dataset from Mirkin et al. (2018) for such claims (iv) a baseline for detecting whether such a claim was mentioned in a speech.

2 Related Work

In Mirkin et al. (2018) we introduced the task of Listening Comprehension over argumentative content. That work analyzes recorded speeches, and tries to identify whether arguments from iDebate[3] are mentioned in the speech. Similarly, in Lavee et al. (2019) we addressed this task by first mining arguments from a large news corpus, and then identifying the arguments which are mentioned in speeches.

This work complements our previous works in two ways. First, the GPR-KB constructed here is of general claims, with wide cross-topic relevance. It facilitates Listening Comprehension for topics not mentioned in iDebate, or topics for which automatic argument mining does not yield satisfactory results. Second, while Mirkin et al. (2018) mention that the iDebate counter points can in principle be used for rebuttal, and Lavee et al. (2019) suggest mining opposing arguments from their corpus to counter arguments mentioned in speeches, pursuing both ideas is left for future work. We pick up the baton (in the context of the GPR-KB suggested here), and annotate the validity of the counter arguments as rebuttal to the ideas

---

[2]Video of the debate at `https://www.youtube.com/watch?v=m3u-1yttrVw`

[3]https://idebate.org/debatabase

expressed in a matching speech.

Response generation has been the subject of much research, using a wide variety of methods (e.g. Ritter et al., 2011; Sugiyama et al., 2013; Shin et al., 2015; Yan et al., 2016; Xing et al., 2017). In the context of dialog systems (see recent survey in Chen et al., 2017), there is usually a distinction between task-oriented systems (Wen et al., 2016) and open-domain ones (Mazaré et al., 2018; Weizenbaum, 1966). The task here can be seen as lying in between the two: on the one hand it allows for a response to speeches on a variety of topics; on the other, the response is restricted to be a rebuttal of a claim made in the speech. A major difference from dialog systems is that in this task the analysis is of a complete speech - rather than taking turns, and the goal is to respond to some of the claims - but not necessarily all.

In the context of computational argumentation much attention has been given to mapping rebuttal or disagreement among arguments. Such works include datasets exemplifying these relations (Walker et al., 2012; Peldszus and Stede, 2015a; Musi et al., 2017), modeling them (Sridhar et al., 2015) and explicitly detecting them (Rosenthal and McKeown, 2015; Peldszus and Stede, 2015b; Wachsmuth et al., 2018). The GPR-KB in this work is reminiscent of argument datasets that depict rebuttal relations, but the arguments are of a different type, being manually authored as general and applicable to a wide range of topics.

Most similar to our work is the task of generating an argument from an opposing stance for a given statement (Hua and Wang, 2018; Hua et al., 2019). These works present a neural-based generative approach, and experiment with user-written posts. Our task differs in that the input is longer text, potentially containing multiple arguments.

3 Data

3.1 Motions and Speeches

The speeches analyzed in this work are the 200 speeches provided by Mirkin et al. (2018). Each speech debates one of 50 motions originating from iDebate. In this data, the phrasing of the motions is often simplified to include an explicit *topic* and *action*. For example, the iDebate motion *This House would introduce goal line technology in football* is simplified to *We should introduce goal line technology*, where the topic is *goal line technology* and the action is *introduce*.

Speeches are evenly distributed between motions, each having two speeches supporting it (i.e. the speaker is arguing in favor of the motion) and two contesting it. They were recorded by 14 different speakers. A speech is given in several formats. We use the recorded audios and manually-created transcriptions. Recordings are about 4 minutes long, and the transcript texts contain on average 28.7 sentences and 833.1 tokens.

Lastly, the dataset contains claims taken from iDebate along with annotations identifying specific claims mentioned in particular speeches. Herein we refer to this data as iDebate18.

3.2 Knowledge base construction

An experienced competitive debater was solicited to author claims that tend to come up in debates across varied topics, and to write a rebuttal argument for each such claim (see the Appendix for the guidelines). She was *not* given access to any of the iDebate18 motions, which are analyzed later on. In total, 39 pairs were constructed in this way.

Texts were allowed to incorporate the special tokens *[ACTION]* and *[TOPIC]*, which are replaced by the debate topic and suggested action when applied to a specific motion or speech. For example, in the context of the motion *We should introduce goal line technology*, the claim *[ACTION] [TOPIC] will encourage better choices* is translated to *introducing goal line technology will encourage better choices*.

In a second phase, the claim-rebuttal pairs were edited by the authors, as follows:

(i) Some rebuttal texts were written with the context of a full speech in mind, and included segments that refer to what a debater would include in such a speech. For example, one included the segment "I have proven that this method is effective". Such segments were edited out.

(ii) For some claims, it seemed that an opposite claim could also be made. In these cases the negation of the claim was also added to the knowledge base, along with an appropriate rebuttal. For example, in addition to the claim "*[ACTION] [TOPIC] is the most practical way to solve the problem.*", we also added the claim "*[ACTION] [TOPIC] is not the most practical way to solve the problem.*".

After these modifications, the final knowledge base includes 55 claim-rebuttal pairs. Claims are always one short single sentence, with an average

Claim	Rebuttal
We must limit personal choice in this case	The greater good means nothing if the rights of individuals are being violated. It doesn't make sense to violate rights in order to protect them.
[ACTION] [TOPIC] is good for the economy	While we need to take the economy into account when making decisions, it cannot be the sole consideration or even the top priority in many cases. In this case, the harms outweigh any benefits there may be to the economy.
We need to protect the weakest members of society	A truly fair society is one where different people are afforded similar rights and are also trusted to look after themselves. While weaker segments of society can be more vulnerable, this does not justify paternalistic policies that are not beneficial for society as a whole.

Table 1: Examples of GP-claims and matching rebuttals, created through the process described in §3.2.

length of 8.5 tokens. Rebuttals are longer, on average 1.8 sentences long, and containing on average 32.9 tokens. Three examples from the GPR-KB are given in Table 1; henceforth we refer to the generated claims as *GP-claims*, or simply as *claims* when the context is clear.

4 Annotation Experiments

Four annotation experiments are described next, aimed at assessing the applicability of the generated GPR-KB to the real world. Each of the following subsections describes one experiment and its results. An overview of the whole process is depicted in Figure 1. The full annotation guidelines for each experiment appear in the Appendix.

4.1 Cross-Topic Relevancy

The GP-claims were written based on the experience of a professional debater, but without context of specific topics. The first annotation experiment aims to establish whether these claims indeed attain the desired goal of being applicable to a varied set of topics. For each motion in iDebate18, and for each GP-claim, we asked annotators to decide whether the claim supports the motion, opposes it or is not relevant[4]. Annotation was done by 7 experienced annotators, and 5 answers were collected for each question.

A GP-claim was considered *relevant* to a motion when marked as supporting or opposing it by most annotators. The stance of relevant claims towards the motion was determined by majority. When a relevant claim has an equal number of supporting and opposing answers, its stance is considered undetermined.
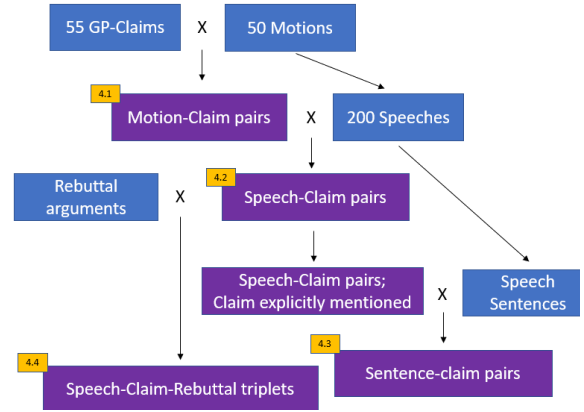

Figure 1: Annotation overview: All motion-claim pairs were annotated for whether the claim is relevant to the motion (see §4.1). For each claim, speeches discussing the relevant motions were annotated for whether the claim was mentioned in the speech (see §4.2), explicitly or implicitly. For explicitly mentioned claims, selected speech sentences were annotated for whether the claim was mentioned in the sentence (see §4.3). In addition, for claims mentioned in the speech, the corresponding Rebuttal Argument was annotated for whether it is a plausible rebuttal in the context of the speech (see §4.4). Blue rectangles indicate textual resources, violet ones indicate annotated resources, yellow ones refer to the relevant subsection.

Results Annotation included 2,750 claim-motion pairs[5], of which 46% are claims annotated as relevant to the motion. 20% are annotated as supporting the motion, 26% as opposing it, and a negligible number have an undetermined stance.

On average, 25.4 claims are annotated as relevant per motion. Inter-annotator agreement (average Cohen's Kappa (Cohen, 1960) over pairs of annotators), is 0.52 for the three-labels task, and 0.45 for the binary label of relevant/irrelevant.

Figure 2 shows the distribution of claims vs. the

---

[4]The stance is required for the experiment in §4.2.

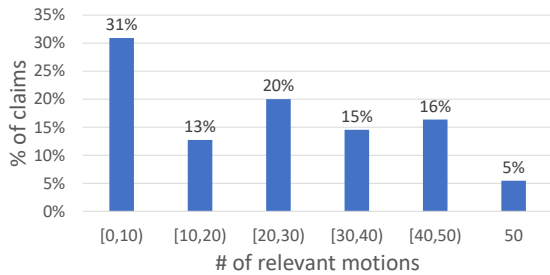[5]All pairs of 50 motions and 55 claims.

Figure 2: The distribution of GP-claims vs. the number of motions annotated as relevant.

number of motions annotated as relevant. Of note, many claims are relevant to various motions: 56% (sum of the four right-most bars) are relevant to at least 20 (out of 50) motions. However, only 5% are phrased in a manner so general that they may be relevant to all 50 motions. An example of such a general claim is *[ACTION] [TOPIC] will harm others*. At the other extreme, the claim *Animals have rights* is labeled as relevant to only 3 motions discussing various animal-related issues: adopting vegetarianism, banning bullfighting and legalizing ivory trade.

The majority of the claims are therefore general enough to be relevant to a substantial portion of the motions, but not so general as to make them trivially relevant to all motions.

4.2 Usage in Spoken Content

Having established that GP-claims are potentially relevant to many motions, the question still remains of whether or not they are actually commonly made by people debating these motions. This is a crucial point for using them as a basis for generating a rebuttal-response.

To assess this, annotators were shown speeches from iDebate18, alongside a matching list of GP-claims determined to be relevant in the previous stage. Specifically, claims annotated as supporting a motion were shown for speeches in which the speaker is arguing in favor of that motion, and vice versa. To allow for a greater number of potential claims, those which at least 2 annotators considered relevant (rather than 3) were included. Claims with an undetermined stance were excluded.

Speeches were presented in both audio and text formats, and annotators were allowed to choose between listening, reading or both. They then had to determine whether each claim is mentioned in

the speech *explicitly*, *implicitly* or not at all. The number of claims presented for each speech was limited to 20; in case a larger number was determined to be relevant, the question was split into chunks of 20 claims.

In total, 3,246 claim-speech pairs required annotation, almost four times more than the corresponding annotation included in iDebate18. Annotation was done using the *Figure-Eight*[6] crowdsourcing platform, with 10 annotators per question. Clearly, this is a challenging task for the crowd, and hence a selected group of 22 annotators was used. Selection was based on their past performance on other tasks done by our team.

To further validate the annotation, the list of claims presented for each speech included claims for which the correct label was known a-priori. These include claims annotated in the previous experiment as irrelevant for the motion, for which it is assumed that the correct label is "not mentioned". In addition, annotation was done in batches. Claim-speech pairs for which unanimous answers were obtained in earlier batches were included in newer ones, with the correct label assumed to be this unanimous answer.

A claim is considered *mentioned* in a speech if a majority labeled it as mentioned (i.e. summing up implicit and explicit answer counts). Otherwise it is considered as *not mentioned*. A mentioned claim is *explicit* in the speech if its explicit answers count strictly exceeds its implicit answers count. Otherwise, it is considered *implicit*.

Results 41% of claim-speech pairs were labeled as mentioned (7% explicit and 34% implicit). On average, each claim is explicit in 4.3 speeches, and implicit in 17.6 speeches.

Pairwise inter-annotator agreement is 0.37 when considering two labels: mentioned or not. The average error rate of all annotators, on questions with a prior known answer, is 7%, suggesting a relatively high-quality annotation.

The *prior* of a claim is defined as the number of speeches in which it was found to be mentioned, divided by the number of speeches in which it was labeled. Figure 3 depicts the percentage of claims vs. their prior, separately for explicit and implicit mentions. Some claims are never mentioned in any speech: 20% are never mentioned explicitly and ~10% are not mentioned at all, not even implicitly. Note that for the most part, claims

---

[6] www.figure-eight.com (formerly CrowdFlower).

are mentioned in less than half of the speeches for which they may be relevant.

In conclusion, the results suggest that GP-claims are often used in spoken content discussing various topics, and that this is not due to a small subset of trivial claims. Rather, most claims appear at least once, but usually in no more than half of the speeches for relevant motions.

These properties make automatic detection of these claims in speeches an interesting and challenging task.

Comparison to iDebate18 Table 2 compares the results of this annotation to that of iDebate18, which contains topic-specific claims annotated for the same set of speeches[7].

Surprisingly, topic-specific claims are no more likely to occur in speeches discussing that particular topic. Moreover, the larger number of potential GP-claims leads to a higher absolute number of mentions, to the extent that – in contrast with iDebate18 – all speeches include at least one mention. Hence, the GPR-KB augments the iDebate18 dataset, both by increasing the number of claims that are to be sought in a speech, and by suggesting claims to speeches for which iDebate18 does not contain any.

Speech stat.	GP-claims	iDebate18
Coverage	100%	86.5%
Avg. Mentions	6.7	1.8
Avg. Potential	16.2	4.4

Table 2: A comparison of GP-claims and topic-specific iDebate claims annotation. Coverage is the percentage of speeches with at least one claim annotated as mentioned. Avg. Potential is the average number of possibly relevant claims per speech, and Avg. Mentions is the average number of claims annotated as mentioned.

4.3 Where was it said?

A straightforward approach to determining whether a claim was mentioned in a speech is to go over its sentences, one by one, and decide whether the claim is equivalent to a sentence or implied by it, as indeed is done in Mirkin et al. (2018). This is a challenging task since, as described above, most mentions are implicit. In many cases one can not point to a single sentence

---

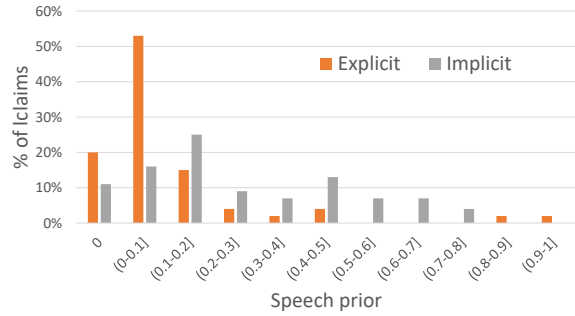[7]Available answers in iDebate18 are *mentioned* or *not mentioned*.

Figure 3: The distribution of GP-claims vs. speech prior (the percentage of labeled speeches in which a claim is mentioned), for explicit or implicit mentions.

mentioning the claim, as the claim is implied by the general stance of the speaker.

Even when a sentence does imply a claim, automatically inferring that may be hard. For example, for the motion *We should end cheerleading*, a relevant opposing claim is *Ending cheerleading limits personal choice*. We identified a sentence implying it, *The only clear moral system we can derive is one in which we value individual preference*, yet it seems hard to deduce this automatically without considering the surrounding context which clarifies that the argument is about personal choice.

An annotation task for identifying *where* a claim was mentioned is considerably more difficult than the aforementioned annotations. Determining ground truth is far from trivial, as annotators may point to different sentences within the same argument as being the location of the claim.

Nonetheless, such information seems a valuable part of a GPR-KB. To provide at least a partial solution, we annotated claim-sentence pairs directly, asking whether the claim is mentioned within the sentence. Algorithms developed on such data can then predict a claim as mentioned in a speech when it is mentioned in one of its sentences. This form of annotation is simple and facilitates easier collection of a large number of labels. To enable research in this direction, such annotations were performed both for GP-claims and, since none are provided in iDebate18, for claims from iDebate.

A careful selection of pairs to annotate is required since there are too many pairs for a comprehensive labeling, and sampling at random would rarely yield a pair such that the claim is mentioned in the sentence. Thus, we limited annotation to claims which were labeled as mentioned explicitly (and assumed all iDebate claims to be

so), and paired them only with sentences which are somewhat similar to them (based on word2vec, Mikolov et al., 2013). Annotation was again done by 10 crowd annotators.

Results Annotation included 4,271 GP-claim and sentence pairs and 2,164 iDebate-claim and sentence pairs, with a similarity of at least 0.5 and 0.7 (resp.) between claim and sentence. The usage of general crowd required some quality control. Annotators not meeting one or more of the following criteria were removed, along with their answers: Answer at least 10 questions; have at least 5 common answers with 3 different peers; have average agreement with peers ≥ 0.2.

The resulting inter-annotator agreement was 0.55 for GP-claims and 0.46 for iDebate claims. Considering only pairs with at least 5 remaining answers, after filtering out annotators as described above, 20% of the GP-claims-sentence pairs were annotated as a match (and 17% of iDebate pairs).

4.4 Validity of Rebuttal Arguments

Recall that rebuttal arguments in the GPR-KB were written without any specific contexts in which they are to be used. Hence, even if the claim they respond to is indeed mentioned in a speech, it is not clear whether the pre-written rebuttal would consist a plausible rebuttal response to the speech.

We assessed the effectiveness of the rebuttal arguments using a two-step procedure. First, as in §4.2, annotators were shown a speech and a claim, and determined whether the claim is mentioned in the speech. Then, if they marked that claim as mentioned, its pre-written rebuttal was shown, and they were asked whether it is a plausible response to the mentioned claim *in the context of the speech*.

This two-step annotation procedure was chosen for three reasons. First, requiring annotators to assess whether a claim is mentioned in a speech motivates them to review its content again and locate the relevant parts in which the claim is expressed. Second, it prevents irrelevant answers from those who do not think that the claim is mentioned in the speech. Finally, asking again about a claim's mention enables result validation, when the answer is known a-priori with high confidence. Specifically, claims for which the annotation is unanimous were used for this purpose.

For each claim, we annotated two randomly sampled speeches mentioning it. This amounted to 103 rebuttal-speech pairs, since not all 55 GP-

claims were mentioned in two speeches. We relied on the same group of crowd annotators who took part in the previous experiment, and once again required 10 answers for each question.

Results Measuring agreement using Cohen's Kappa in this scenario is problematic. First, the label distribution is very biased: If rebuttal arguments are always plausible responses, then the correct answer is always yes. Any deviation from that will greatly reduce the score, making it an ill-fitting measure for such data (Jeni et al., 2013). Second, the small number of questions leads to many annotator-pairs whose set of common questions, on which this score is computed, is rather small. This makes the computation unstable, since when averaging over all pairs, such small intersections contribute as much as the large ones.

Instead, looking at the majority annotation for each rebuttal argument, we observed that in 87% of the cases it indicates that the rebuttal is plausible. This suggests that regardless of whether annotators agree with one another, they tend to agree that the rebuttal is usually plausible. Moreover, computing the average kappa between annotators and the majority annotation, and considering only those annotators who answered at least 20 questions, yields 0.47. This value for such biased data, alongside an average error rate of 4% on the questions with known answers, suggests that the annotation is of reasonable quality.

Analysis The results above show most rebuttals are appropriate in the vast majority of contexts. We therefore decided that continuing with this costly annotation is not needed. Furthermore, manual analysis of cases in which the rebuttal was unanimously found inappropriate showed this stemmed from the rebuttal being inappropriate for the topic, rather than a specific speech. For example, when discussing goal-line technology, in response to the claim *"introducing goal-line technology will lead to greater problems in the future"*, the pre-written rebuttal is *"Governments have an obligation to their citizens in the here and now. The better off society is today, the more resources we will have to make the future better when it comes"*. Such a response makes several assumptions which are clearly violated here, such as the involvement of government. Thus, further validation of rebuttals may benefit from first verifying their relevancy to the topic.

4.5 The GPR-KB-55 Dataset

The annotation results show that it is possible to construct a concise set of general claims, such that in most speeches at least one of them will come up. Furthermore, they show that a rebuttal to these claims can be authored independently of the specific motions and speeches, while nonetheless being a plausible response in their context. Table 3 summarizes the statistics for the pair-annotation experiments. The resulting dataset is freely avail-

Annotation	# pairs	# positives
Motion – GP-claim	2,750	1,265
Speech – GP-claim	3,246	1,491
Sentence – GP-claim	4,271	854
Sentence – iDebate claim	2,164	368

Table 3: Summary of annotation experiment results. Positive pairs are those in which a majority of the annotators indicated that the claim is relevant (motion) or mentioned (speech, sentence).

able, and is one of the main contributions of this work[8]. We name the new dataset GPR-KB-55.

5 Detecting claims in speeches

Next, we establish baseline results for determining whether a GP-claim is mentioned in a speech, and compare them to results obtained for iDebate claims. For a fair comparison of the two data sources, we assume for both prior knowledge as to which claims are relevant for a motion, as well as their stance towards it. Hence, we take the labeling of GP-claims to motions (described in §4.1) as given. The following algorithms are considered:

word2vec The best performing baseline of Mirkin et al. (2018) utilizes a detailed description of each iDebate claim, comprised of several sentences. It examines the speech sentence by sentence, and for each sentence computes its tf-idf weighted word2vec (*w2v*) similarity to the detailed claim description. A claim is then scored by taking the maximum over all claim-sentence similarity scores. We use this method (*w2v-i18*) as a baseline for the GP-claims as well, yet sentences are scored by their similarity to the GP-claim text, since no detailed topic-specific description exists. The latter is referred to as *w2v-GP-claims*. For

comparison, we repeat the experiment using only the iDebate claim texts (*w2v-i18-claims*).

Bert Recently, the Bert architecture (Devlin et al., 2018) has proven successful on similar tasks, and we provide its results as an additional baseline. Specifically, we select at random 80% of the motions as an ad-hoc train set, and the remainder as a test set (*bert-test*). Bert was trained on labeled claim-sentence pairs corresponding to motions from the train set, in two settings, considering: (i) GP-claims ($\sim$3K pairs) – *bert*, and (ii) both GP-claims and iDebate claims (almost 5K pairs) – *bert+*. In inference, given a claim and a speech, sentences semantically similar to the claim (as in §4.3) are scored by the fine-tuned network. Their maximum is the outputted claim-speech score.

Prior One important difference between GP-claims and iDebate claims is that the same GP-claim can (and does) appear across different motions and speeches. Specifically, given a training set, the a-priori probability that a GP-claim will be mentioned in a speech can be computed. Then, test claims are scored with their computed a-priori probability *without considering the text of the speech*. This baseline is referred to as *prior*.

Results Figure 4 plots precision-recall curves comparing claim detection baselines over iDebate claims and GP-claims. As observed by Mirkin et al. (2018), *w2v* works best when given a detailed iDebate claim description. Without it, performance is comparable for the two claim sources, and is rather poor for both. *Prior* results were obtained by using a leave-one-motion-out cross validation: at each fold a single motion is left out and the others are used for training. Its precision-recall curve shows that when considering these statistics, it presents a challenging baseline.

To compare the *bert* baseline to others, the precision-recall curves for both *prior*[9] and *w2v* were computed over speeches from *bert-test*. As shown in Figure 5, while *bert* clearly outperforms *w2v*, it nonetheless does not reach the *prior* baseline. The additional data provided in training to *bert+* does not help.

Note that this comparison is between methods which are derived from different types of data. Here *bert* is trained only on explicitly-mentioned claims, with respect to (ostensibly) semantically

---

[9]Here, a-priori probabilities were computed on all motions not in *bert-test*, then applied to motions of *bert-test*.
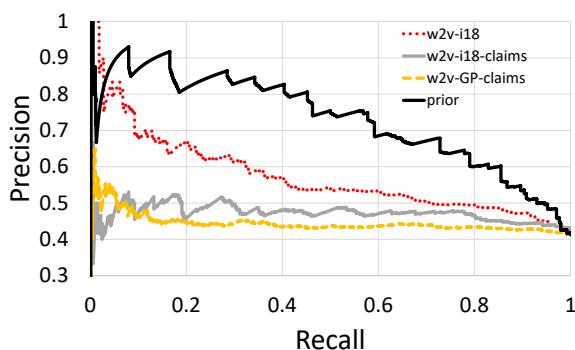
Figure 4: Precision-Recall curves for the matching of GP-claims and iDebate claims to all 200 speeches.
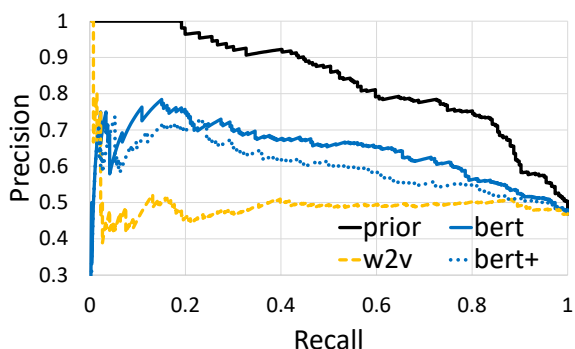

Figure 5: Precision-Recall curves for matching GP-claims to speeches, for *bert-test* (20% of motions).

similar *sentences*. On the other hand, the *prior* baseline is computed based on all claims, and their annotation w.r.t. *the entire speech*. This may be part of the reason why *bert*, which has proven to be successful on many NLP tasks, here achieves lower performance than this simple baseline.

Analysis Although *prior* seems like a strong baseline in terms of precision and recall, it is probably not a desired solution by itself, since it simply produces high probability responses regardless of the rebutted content. For example, among its top-20% predictions, precision is 83% and recall is 40%, yet they include only 22% of available GP-claims. Moreover, 77% of these top-20% are the same 5 claims. This reflects a property of the data: there are a few claims which are relevant to many motions, and are also implicitly mentioned in most speeches. Detection algorithms should be aware of this property, and account for it when evaluating performance. At the same time, a claim's acceptance prior can be useful for inference. For example, it could be combined with other data in a more sophisticated algorithm, or could direct the parameter choice of such an algorithm.

6 Conclusions and Future Work

We presented the problem of producing a rebuttal response to a long argumentative text. This task is especially challenging when the discussed topic is not known in advance, and, accordingly, potential responses are not readily available.

Toward the goal of addressing this problem we constructed a multi-layered dataset: (i) A Knowledge base of GP-claims and corresponding rebuttal arguments, which are shown to be applicable for a wide variety of topics; (ii) A mapping of these claims to motions of iDebate18 in which they might be applicable; (iii) An annotation of the stance of applicable claims; (iv) An annotation for which claims are actually mentioned in relevant speeches, and whether they are mentioned explicitly or implicitly; (v) For explicitly mentioned claims, a (partial) annotation of which sentences imply them and which do not.

In addition, we presented baselines for the related Listening Comprehension task, suggesting that this is a complicated problem. Using state-of-the-art sentence embedding yielded an $F1$-score of 0.64, while trivially taking the claim with the highest prior to be mentioned scored 0.78[10]. This suggests that careful evaluation is required.

While baselines are provided only for detecting GP-claims in spoken content, future work should aim to solve the problem as a whole - either by developing algorithms that determine relevance and stance of GP-claims to given motions, or by forgoing these stages, and successfully deciding whether a claim was mentioned in a speech, without first focusing on relevant claims.

Our results suggest that GP rebuttal arguments usually work well as a response to speeches in which the matching claim was mentioned. However, this is by no means perfect; and in some 13% of the cases they do not. It is interesting to further identify and understand these cases. By doing so, an automatic system could prefer responses that it identifies as more appropriate. Moreover, understanding such cases can lead to improving the rebuttal arguments themselves.

7 Acknowledgments

---

[10]Scores computed at the precision-recall intersection.

References

Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. 2019. Argument invention from first principles. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1013–1026, Florence, Italy. Association for Computational Linguistics.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

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.

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.

László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251. IEEE.

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

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *EMNLP*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Shachar Mirkin, Guy Moshkowich, Matan Orbach, Lili Kotlerman, Yoav Kantor, Tamar Lavee, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2018. 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)*.

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

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

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

Sara Rosenthal and Kathy McKeown. 2015. I couldnt 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.

Andrew Shin, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2015. Context-dependent automatic response generation using statistical machine translation techniques. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1345–1350.

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.

Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. 2013. Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures. In *Proceedings of the SIGDIAL 2013 Conference*, pages 334–338.

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.

Joseph Weizenbaum. 1966. Eliza–a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve J. Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. In *EACL*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64. ACM.