nature portfolio

https://doi.org/10.1038/s41586-021-03215-w

Supplementary information

An autonomous debating system

In the format provided by the authors and unedited

An Autonomous Debating System - Supplementary Materials

Contents

1 Debate Format

The debate format we considered is a variation on British Parliamentary Style Debates. Specifically, the debate is between two sides, referred to as *government* and *opposition*, with the government being represented by Project Debater, and the opposition by a human debater. The debate proceeds in three rounds, where in each round Project Debater speaks first, followed by the human debater. The speeches of the first and second rounds are each limited at four minutes, while those of the third round are at most two minutes long. Roughly speaking, the goal of the first round is to present the main points of each side; that of the second round is to rebut the other side's points and develop your own; and that of the final round is to suggest additional rebuttal, and summarize your own points and the debate in general.

Both sides have 15 minutes to process the motion and prepare for the debate. While the human debater prepares right before the debate, Project Debater processes the motion beforehand but takes no longer than 15 minutes to do so. Prior to the actual debate, the moderator introduces the motion, and outlines the main controversies associated with it. Following the format of Intelligence Squared US, a radio program and podcast associated with NPR, the audience of the debate votes on the motion both before the debate starts (after hearing the moderator) and after it concludes. In both cases one can be for the motion, against it, or undecided. The side for which support increased by a greater fraction of the audience is defined as the winner of the debate. In our case, the audience were further asked to vote which side better enriched their knowledge about the debate topic.

2 Motion Phrasing

In general, a debate's motion can range in scope from touching upon a high-level social or economic issue, e.g., This House would unleash the free market, to being very specific, discussing a particular case, e.g., This House believes Tennessee is correct to protect teachers who wish to explore the merits of creationism (both examples are taken from Debatabase, http://idebate.org/debatabase/index.php).

Early in the project, it became clear that for a system to successfully retrieve a sizable number of arguments – even from a very large corpus – the motion should not be very specific. That is, while there are probably some arguments in our corpus relevant to the motion in the second example, there are probably not enough of them for a precision-oriented system to extract. Moreover, a very specific debate might be of limited interest to a general audience and might also be inappropriate for human debaters who are not familiar with it. More precisely, an experienced debater will usually turn the debate from dealing with the specifics to the more general principles underlying the discussion, and hence it makes sense to phrase the debate in this way from the onset.

Specifically, we focused our attention on motions conferring to one of the following formats: (i) Policy motions, of the format: We should [**action**] [**topic**] or [**topic**] should be [**action**].

(ii) Analysis motions, of the format: [**topic**] [**view**].

The topic of the motion is defined as any title of a Wikipedia article, or one of its redirects. The rationale is that if a **topic** is interesting enough to be debated, then it is very likely to have an associated Wikipedia page. We ensured that the motion sets employed in our evaluations do not include duplicate topics (for different actions/views), nor near duplicates which are likely to imply essentially the same debate. In addition, we ensured that the different sets used for training and evaluation are distinct in terms of topics (also not allowing near duplicates in different sets).

The **action** or **view** of a motion comes from a pre-defined list of terms, which all have a clear sentiment towards the **topic**. These include actions that define proposed policies, such as 'ban', 'legalize', 'adopt', 'increase the use of', 'subsidize', etc.; and views that reflect an opinion, such as 'brings more harm than good', or 'is justified'.

In addition, we required that the motion phrasing would not imply a debate focused on the validity of facts, but rather allow for controversy stemming from different point of views and principles.

An example of a policy motion phrasing that follows the above principles is *We should subsidize preschool* (alternatively, *Preschool should be subsidized*), where the **topic** is “Preschool”, and the **action** is “subsidize”. An example of a legitimate analysis motion phrasing is *Al brings more harm than good*, where the **topic** is “AI” (which redirects to “Artificial Intelligence”), and the **view** is “brings more harm than good”.

Finally, note, that the motion can be phrased with the opposite stance, e.g., “We should *not* subsidize preschool”, resulting with Project Debater in side government, expected to *oppose* the notion of subsidizing preschool.

3 Motion Collection

Many of the underlying components of Project Debater – and the argument-mining components in particular - are trained via the classical supervised learning paradigm. Correspondingly, to achieve high performance, these components typically require a large and representative training set, composed of carefully labeled examples (sentences pertaining to the debate), that ideally correspond to a large and diverse set of motions.

There are several publicly available collections of debate motions. Originally, we considered motions from Debatabase (https://idebate.org/debatabase), that were easily phrased using the motion phrasing guidelines described above. Since this gave rise to a relatively small set of motions, we further defined novel motions using the following scheme.

Controversial Wikipedia titles were identified by automatic means [1], followed by manual annotation by in-house annotators and crowd workers. Next, in-house annotators were presented with topics deemed controversial in the previous task and asked to attach to them the most appropriate **action** or **view**. A team of three expert annotators reviewed the resultant motions and discarded those which did not confer to the motion phrasing guidelines described above. This process gave rise to the motion sets, described next.

4 Motion Sets

When evaluating or demonstrating the system in a live debate, it is of course crucial that the motion would be selected from a set of motions that were never included in the training/development data of any of the underlying components of the system.

To that end, we have set aside a collection of motions, termed henceforth the Unseen Set, which were not available for the routine development and assessment of the project, and from which motions were selected for live events.

The motion collection process described above was performed in batches. From each batch, a few of the new motions were set aside for the Unseen Set – excluding motions that may give rise to offensive arguments, and hence could not be used in live events (e.g., We should ban pornography), while the rest were used for developing and evaluating the system as described next.

Towards the conclusion of the project we already accumulated nearly 700 motions, beyond those assigned to the Unseen Set, that were used differently by the different components of the system. The Argument Knowledge Base component, which matches motions to Classes of Principled Arguments, used all these 700 motions to train its classifiers. This was necessary since the basic unit being classified is a motion, and even for simple classifiers, at least several hundred examples are typically needed for effective training. Evaluation of the classifiers was done using a leave-one-out framework, with classifier thresholds set accordingly.

Other components, such as the argument-mining classifiers, were typically developed to classify sentences, hence it was possible to train and develop these components over a smaller – yet representative - sample of motions. Correspondingly, we defined the following subsets of motions, all not from the Unseen Set:

Component-development set – used to develop each component independently of the others. This set eventually included 100 motions used for training and development of neural models used for the argument-mining tasks, and 60 additional motions that were further used for setting classifiers' thresholds and evaluating the performance of individual components on a regular basis, in a leave-one-out framework.

Pipeline-set-1 – This set, which eventually included 78 motions, was used to assess the system in its entirety, including the complete argument mining pipeline which combines the components developed using the component-development set. That is, argument-mining classifiers were developed and trained on the component-development set of motions, while their joint performance in the complete pipeline was evaluated on this Pipeline-set-1. Essentially, Pipeline-set-1 was serving as a test-set of the entire system. However, continuous usage of a fixed test set while developing a complex multi-component system, may result with gradually overfitting the overall system behavior to the characteristics of this set, in a manner that will be hard to assess or to control. To address this concern, we established a second pipeline-set, that was defined at a much later stage of the project and is described next.

Pipeline-set-2 – This set included 158 motions, selected during 2018, and was also used to assess the system performance in its entirety. The need for this additional set arose from concern about overfitting, as described above, but also from a need to better represent the Unseen Set. Since motion collection was a gradual process, it was only at this late stage in the project that we could more clearly define the general statistics of the Unseen Set. The motions in Pipeline-set-2 were selected with the aim of better capturing these statistics, mainly the distribution of action terms in the motions of the Unseen Set.

Out of Pipeline-set-2 we sampled 36 motions for the analysis described in this paper. Accordingly, we also selected 36 motions from Pipeline-set-1, with a similar distribution of motions' **action**. We denote these subsets *Eval-2* and *Eval-1*, respectively.

5 Project Debater's Components

In the following sections, we detail the different components of which Project Debater is composed, the tasks they aim to solve, and how they interact which each other. Although each component is built to pursue a separate task, we have also developed over the years cross-component text analysis tools, which are task-independent:

Wikification. The goal of our Wikification tool [2] is to identify *mentions* – phrases in a given text that are mapped to Wikipedia concepts (titles). We refer to this tool henceforth as *Wikifier*.

Semantic Relatedness of Wikipedia Concepts. The goal of our semantic relatedness model [3] is to assess the similarity between two Wikipedia concepts. E.g., between two concepts previously extracted from a given text by the Wikifier. We refer to this model henceforth as *WORT* (Wikipedia Oriented Relatedness Tool).

For Argument Mining we rely on a large corpus of some 400 million newspaper articles[1]. We process this corpus, breaking the articles into sentences, and indexing these sentences not only by the surface forms of the words therein, but also by the Wikipedia concepts they refer to, the entities they mention [4, 5], pre-defined lexicon words [6], and so on. Beside argument mining, this index is also utilized by other components, whenever some corpus-wide retrieval is required (e.g., for collecting training data to train a neural model in Section 5.1.4). We refer to this indexed corpus henceforth as *LexisNexis*.

5.1 Argument Mining

5.1.1 Claim Detection

The first core task of the Argument Mining component is that of Claim detection. This task, initially described in [7], defines a claim as a concise statement which has a clear stance towards the motion (see also Section 8). Claims are one of the primary sources of arguments, alongside evidence, which are mined from the LexisNexis corpus for speech content. The claim detection pipeline, described next, consists of retrieving sentence candidates from the LexisNexis corpus; ranking them for their likelihood to contain a claim; extracting from each sentence an exact boundary which is most likely to match a claim; and finally, ranking all extracted candidate

[1] From the LexisNexis 2011-2018 corpus, `https://www.lexisnexis.com/en-us/home.page`

claims. The general outline of the pipeline is presented in [7], though the practical details differ from what we describe there. The main differences are that we retrieve sentences by querying a sentence-level index, instead of extracting them from retrieved articles; and that the sentence ranking mechanism is based here on a neural model.

Sentence retrieval. Candidate sentences are retrieved by several queries which are designed to identify patterns of sentences which frequently contain claims. These queries were developed by considering only the motions in the Component-development set (See Section 4). The core principles of the queries can also be found in [8, 6]. The goal of this step is to identify sentences which are likely *claim sentences*, that is, sentences which are likely to contain a relevant claim. Below are the building blocks used to build the queries:

1. **"that"** - the term "that" has been shown to be a prominent feature indicating that a sentence contains a claim [7].

2. **Query sentiment lexicon** - a subset of the expanded Hu and Liu lexicon [9] described in detail in Section 5.1.3, trimmed to contain only words commonly found in sentences containing claims. The full lexicon is available in the supplementary files.

3. **Action verb expansions** - syntactic and semantic expansions of common action verbs in debate motions. For example, expansions of the action *ban* include *banned*, *bans*, the synonym *outlaw* and the antonym *legalize*. The full list is available in the supplementary files.

4. **Claim verb phrases** - verb phrases commonly found in claim sentences. For example, *leads to*, *causes* and *is harmful*. The full list is available in the supplementary files.

Next, we describe the set of queries used to retrieve sentence candidates. The terms listed in each query need to appear within the sentence in the listed order, but need not be adjacent within the retrieved sentence - yet all terms must appear within a window of 12 tokens.

1. “that” + **topic** + Sentiment
2. “that” + **topic**
3. Action verb expansions + **topic** + Verb phrase
4. **topic** + Action verb expansions + Verb phrase
5. “that” + Action verb expansions + **topic** + Sentiment
6. “that” + **topic** + Action verb expansions + Sentiment
7. **topic**

All sentence candidates are constrained to contain between 5 and 50 tokens. Queries 1-2 are run together in a cascade – query 2 is executed only if query 1 did not retrieve the requested amount of sentence candidates. Queries 3-6 are executed separately in a similar way, and their results are appended. Query 7 is run independently from the others. Its much broader search aims at retrieving claims which may have been missed by the previous, more specific queries. Accordingly, it is set to retrieve a greater number of sentence candidates, as the a-priori ratio of sentences containing claims retrieved by this query is presumably much smaller.

Sentence ranking. All retrieved sentences are ranked by a neural model for their likelihood to contain a claim. Our network is a bi-directional LSTM with an additional attention layer [10], described fully in [6]. For more details about the process of collecting labeled data, see Section 8. Our labeling task focuses on statements marked within their respective sentence (a claim *boundary*). For the purpose of training the sentence ranking model, we consider sentences containing accepted statements (claims) as positives, and sentences containing rejected statements (non-claims) as negatives.

Claim boundaries detection. For each of the top-scoring sentences, we extract the most probable claim boundary by a Maximum Likelihood probabilistic model, described in [7], referred to there as **Boundaries Coarse Filter**. The reason for employing a simple statistical method at this stage is that the number of boundary candidates to consider is relatively high, and could thus negatively impact the run time if processed by a more complex classification model. The probability is calculated by multiplying the probabilities of the word which precedes the boundary, the word which is immediately after it, the part of speech of its first token, and the part of speech of its last token. For each sentence, we traverse the top-50 boundaries in descending order and apply several filters, which relate to the structure of the text within the boundary (e.g., contains a verb without its subject, contains a newline) or to its content (e.g., does not contain the **topic**, starts with a cohesive marker, is too long or too short). Following the filtering process, for each sentence we keep the first boundary which passes all filters successfully. After this boundary is selected, we apply an additional modifier, where we use several lexical patterns to identify if the sentence expresses a negative stance towards the selected boundary, e.g., *I do not think that [gambling is addictive].* In this case, we replace the selected boundary with the boundary of the entire sentence. The motivation for this is that we want the claims in the speech to represent the true intent of their source.

Claim ranking. Finally, we apply a logistic regression classifier for ranking all selected claims, described in [7]. The features used to model the classifier include lexical, semantic and syntactic input from the claim, the containing sentence, and the motion. For example, whether the sentence's ESG parse tree [11] contains a subordinate "that" whose sub-tree matches the claim; the ratio of sentiment words; and the identity of parts of speech right before, at the start, at the end, and right after the claim. Finally, the claim score is set to be the average of the Claim ranking and Sentence ranking scores, and the top claims are passed on to downstream components.

5.1.2 Evidence Detection

While a claim is a concise (usually sub-sentential) statement supporting or contesting the motion, an evidence is a full sentence (or longer) which provides information aiming to convince the audience to accept the presented stance towards the motion (or reject it, in the case of a counter evidence). For more on the definition of evidence, see Section 8.

After considering several evidence types [12], we focus on detecting two types: (i) *Study* evidence, which presents results of a quantitative analysis of data, or conclusions of a research, and (ii) *Expert* evidence, which presents a testimony of an authoritative figure with some known expertise on the **topic** (e.g., a person, a committee, an organization). We chose these evidence types based on their potential to contribute to a fact-based debate. In addition, the ability to pinpoint supporting pieces of information in Big Data is a core strength of an AI system. It complements human capabilities of appealing to the emotion of the audience, by presenting other evidence types, such as an anecdotal story.

Each evidence type has its own pipeline, detailed next, which shares some resemblance to that described in Section 5.1.1. Namely, it consists of retrieving candidates from the LexisNexis index; ranking them by their likelihood to represent valid evidence; and finally, identifying in the top of the candidate list, those of a higher quality.

Sentence retrieval. Similarly to Section 5.1.1, candidates are retrieved by several queries which are designed to cover elements that are frequently presented in evidence texts. We define the following categories which are used as building blocks of our queries:

1. **Study mention** - synonyms and alternative terms for mentioning a study: *study*, *research*, *report*, *poll*, *survey*, *analysis*, *observation*, *exploration*, *documentation*, *meta analysis*.

2. **Expert mention** - terms that are used to describe an authoritative figure worth quoting: *expert*, *doctor*, *scientist*, *researcher*, *faculty*, *economist*, *philosopher*, *chief*, *commission*, *scholar*, *investigator*, *council*, *professor*, *guru*, *specialist*, *analyst*, *critic*.

3. **Numeric expression** - terms identified as *Number* by Stanford NER [13], or as *Currency* or *Percent* according to their DBPedia entry.

4. **Query sentiment lexicon** - see Section 5.1.1.

5. **Action verb expansions** - see Section 5.1.1.

6. **Study conclusion** - phrases (unigrams to 5-grams) that frequently appear in reports of a study result and its conclusions. It contains phrases like *has a positive effect on*, *is associated with*, and *may reduce the risk of*. The full list is available in the supplementary files.

7. **Person** - terms identified as *Person* by Stanford NER or annotated as such in DBPedia.

8. **Organization** - terms identified as *Organiztion* by Stanford NER.

Next we describe the set of queries we designed for each evidence type. The queries differ in their precision and coverage, and aim to retrieve a diverse set of candidates for the rest of the pipeline. As with claims, terms need not be adjacent.

Expert queries:

1. Person + Expert mention + **topic**

2. (Person or Organization) + “that” + **topic** + Sentiment

3. (Person or Organization) + “that” + **topic** + Study conclusion

Study queries:

1. Action verb expansions + **topic**
2. **topic** + Numeric expression + (Study conclusion or Sentiment)
3. Study mention + “that” + **topic**
4. Study mention + “that” + **topic** + Sentiment
5. Study mention + “that” + **topic** + Study conclusion
6. **topic** + (Study conclusion or Sentiment)

After retrieval, several filters are applied to filter out mal-constructed sentences (e.g., not well parenthesized, contain a newline mark), or improper candidates for evidence (e.g., are too long, start with a pronoun, end with a question mark, contain a contradiction).

Sentence ranking. All sentences which pass the filters are then scored by a neural model for their likelihood to be a valid evidence. The architecture of the network used here is similar to that used for Sentence ranking in Section 5.1.1, described also in [14]. In a later work, after the live debate event [15], we replaced this network with BERT [16] and further improved evidence detection performance. For more details about the process of collecting labeled data, see Section 8.

The top scoring candidates are filtered by two filters whose run time is more demanding than the simpler filters applied after retrieval, and are thus placed at this stage of the pipeline so they are used to process only a limited amount of candidates. The first filter removes candidates which mention the name of person who does not have an entry in Wikipedia. The rational is that people which do not have a Wikipedia page are probably not authoritative figures and we better not consider them as experts. The second filter identifies sentences with unresolved anaphora as we want to avoid using such context-dependent sentences in our speech.

Evidence quality. Since only a few evidence sentences will eventually be selected to the speeches, we want to give preference to the more effective ones. To that end, we explored the notion of argument quality, which is gradually receiving more attention from the argument mining community (e.g., [17, 18]).

We started with creating a dataset of evidence pairs where each pair is manually annotated to indicate which evidence is the more convincing one out of the two. With this dataset we trained a Siamese neural network to identify the more convincing evidence candidate and to reorder the top candidates list accordingly. The creation of the dataset (which we made publicly available) and the details of the Siamese network are described in [19].

Another method we have implemented in order to provide evidence of a higher quality was applied only to the Study type. For every Study evidence sentence we examine its subsequent sentence. If it contains additional details about the quantitative results of the study, we merge the two sentences to obtain a more informative and complete evidence. Such elaborated Study evidence candidates receive a higher priority to be selected for the speech, but where limited to be used only once during a debate. We identify a sentence as presenting detailed quantitative result if one the following applies: (i) it contains the Percent or Money named entities (identified by Stanford NER); (ii) it contains another type of number (except a year) which is immediately preceded with one of the following prepositions: [*about*, *almost*, *around*, *as far as*, *by*, *from*, *of*, *between*, *for*, *up to*, *exceeding*, *than*]; (iii) it contains a number and a term which indicates a monetary value or percent: [*%*, *percent*, *$*, *dollar*, *USD*, *EUR/s*, *euro/s*, *million*, *billion*, *per*, *a year*, *each year*, *annually*, *a month*, *each month*, *monthly*]. These manually written rules were validated with a corresponding analysis.

5.1.3 Stance Detection

The goal of the stance detection task is to identify the stance of claims and evidence towards the motion. We use the stance prediction model first described in [20] as follows. Given a motion m and an argument a, let x_m and x_a be their sentiment targets, respectively, and let $s_m, s_a \in [-1, 1]$ be the sentiment in the motion and argument toward those targets. Then let $\mathcal{R}(x_a, x_m) \in [-1, 1]$ denote the relation between the argument target and the motion target where positive values indicate a consistent relationship and negative values a contrasting relationship. Combined, the stance of a toward m is defined as: $Stance(a, m) = s_m \times \mathcal{R}(x_a, x_m) \times s_a$ with positive values indicating that a is supporting the motion (a *pro* stance) and negative values indicating that it is opposing the motion (*con*). For example, if the motion is *"We should ban smoking"*, and the argument is *"Smoking causes cancer"*, the argument supports the motion because they both have negative sentiment towards their consistent targets. In this case, $x_m = x_a =$ *smoking*, $s_m = s_a = -1$ and $\mathcal{R}(x_a, x_m) = 1$, and $Stance(a, m) = (-1) \times 1 \times (-1) = 1$.

We aimed to further decompose arguments of type *evidence* as follows. These arguments often contain an inner claim, and may express a negative stance towards that claim, which flips the overall polarity. This is similar to the identification of a negative stance of a sentence towards its claim, described in Section 5.1.1. For example, given the evidence *"A study conducted in Germany by von Salisch found* ***no evidence*** *that violent games caused aggression in minors"*, we extract the inner claim *"violent games caused aggression in minors"*, and the negative stance towards it, expressed by *"no evidence"*. We found that a relatively small set of patterns is sufficient for this task. The stance of the inner claim towards *violent games* is then computed using the method detailed below, and the resulting negative stance is flipped due to the negative stance towards the claim, yielding an overall positive stance for the argument.

For the purpose of detecting stance, we assume that the motion target x_m and sentiment s_m are given. The argument target x_a and the relation $\mathcal{R}(x_a, x_m)$ are extracted simultaneously. Each argument contains the **topic**, or an expansion of it (see Section 5.1.4). We refer to this mention as the *debate concept mention (DCM)*. If a noun phrase subsuming the DCM contains a term from a lexicon of contrastive expressions (provided in the supplementary files), the whole noun phrase is extracted as the argument target, and the relation is set to -1; otherwise, the argument target is the DCM alone and the relation is 1. For example, if x_m = *"smoking"* and the argument is *"Smoking ban is too difficult to enforce"*, then x_a=*"Smoking ban"*, and $\mathcal{R}(x_a, x_m) = -1$.

Our main challenge was detecting s_a, the argument sentiment towards its target. Over the course of the project we have developed different sentiment analysis approaches for arguments, and for the final system we have empirically chosen the approach best suited for each argument type. Below we describe two systems for sentiment analysis: a knowledge-based system and a neural-based system. Both systems can be used for both claims and evidence, but in the final Project Debater configuration we used the neural-based system for predicting the stance of claims, and a hybrid approach for the more challenging task of evidence stance prediction.

Knowledge-based system. The knowledge-based system combines components that leverage lexical-semantic or world knowledge. The most important component performs sentiment analysis combining sentiment lexicons, polarity shifter lexicons, and sentiment weighting in relation to the target (see [20]).

We have developed several sentiment lexicons, which were used to detect sentiment phrases in the sentence. We started with the publicly-available sentiment lexicon of Hu and Liu [9], comprising a few thousands of terms. Using this lexicon, we trained an SVM classifier for predicting word sentiment from its embedding, which allowed us to substantially expand the initial lexicon [21, 22]. The resulting lexicon was filtered based on WordNet [23] relations as well as crowd annotations. We have also annotated for sentiment a lexicon of $5{,}000$ common idioms, e.g., *"under fire"* and *"make a difference"*, resulting in the largest available resource of its kind [24].

Additional lexicons were used for addressing various sentiment composition phenomena, such as *valence shifters* (*"hardly helpful"*) and mixed-sentiment phrases (*"cure cancer", "sophisticated crime"*). These lexicons were developed in a semi-automatic fashion: we developed a fully-automatic method for extracting lexicons for sentiment composition [22], which were then manually reviewed and expanded.

For evidence of type Expert, the above sentiment-based method was combined with a complementary approach, which utilizes information external to the argument to determine the expert's stance towards the **topic** [25]. For example, knowing that Richard Dawkins is a famous atheist gives us a strong prior about the stance of his arguments on atheism. We considered two sources of information: (i) Wikipedia categories (Dawkins, for instance, is a member of the *Atheists* category), and (ii) sentences in our LexisNexis corpus that contain both the expert and the **topic**, for which we train an SVM classifier that aims to predict the expert's stance from the contexts these sentences provide.

After these components are run, their output and other information extracted from the argument are used as features for a reranker, implemented as an SVM classifier. The goal of this component is to rank higher predictions that are likely to be correct. Its features model argument complexity, confidence of stance sub-components and their disagreements.

Neural-based system. As discussed above, the neural component is trained to perform targeted sentiment analysis, i.e., it predicts the sentiment s_a of the argument a towards its target x_a, which is masked in the input sentence. The basic network architecture is an RNN using GRUs

[26], which feeds into an attention layer [10] and then a dense layer with ReLU activation and finally a softmax layer for the output prediction. The RNN takes a word2vec [27] word embedding representation for each word in the argument. We further augmented this standard RNN to incorporate features from the knowledge-based system. First, each token's word embedding features were appended with features capturing the token's sentiment, importance, and whether it has been negated or the valency has shifted. Then we add sentence-level features from the knowledge-based system (sentiment scores, and counts of sentiment words and shifters), which are fed into a dense layer with ReLU activation, and the output of this is concatenated with the RNN output before going to the softmax layer. The softmax output is then scaled to $[-1, 1]$ to get the final s_a. The network is trained on sentiment-labeled claim and evidence arguments (the whole evidence, or its inner claim, if extracted).

As mentioned above, for evidence the final prediction combines neural and knowledge-based predictions. A simple but effective approach for high-confidence prediction is only taking arguments where both approaches agree, and otherwise we consider the argument's stance as neutral and ignore it.

5.1.4 Topic Expansion

When debating a controversial topic, it is often desirable to extend the boundaries of the discussion, and bring up arguments about related topics. For example, when discussing the pros and cons of the *presidential system*, it is natural to contrast it with those of the *parliamentary system*. When debating *alternative medicine*, we may discuss specific examples, such as *homeopathy* and *naturopathy*. Conversely, when discussing *bitcoins*, we can speak more broadly on *cryptocurrency*.

Debate Topic Expansion [28] aims to find related topics that can enrich our arguments and strengthen our case when debating a given topic. Similarly to the **topic**, expansions are

Wikipedia titles as well. We distinguish between two types of expansions: *consistent* and *contrastive* [20]. Arguing in favor or against a consistent expansion may support the *same* stance towards the **topic**, whereas for contrastive expansions the stance is reversed. For example, *Bitcoin⇒Cryptocurrency* and *Alternative medicine⇒Homeopathy* are consistent expansions, while *Presidential system⇒Parliamentary system* is a contrastive expansion, since we may support the presidential system by criticizing the parliamentary system. Topic expansion comprises the following steps:

Candidate extraction. First, candidates are extracted from a large corpus using a set of predefined patterns. Each expansion type makes use of a different type of corpus and patterns. For example, patterns for consistent expansions include 'X *such as* Y' and 'X *and other* Y', which aim to extract expansions that are either broader or more specific. This type of expansion is extracted from our LexisNexis corpus. One of the slots X, Y should match the **topic**, and the other slot should match its extracted expansion. Patterns for contrastive expansions denote comparison or alternatives, e.g., 'X * *vs* * Y', and '*difference between* * X * *and* * Y', where '*' matches any number of non-concept tokens. Contrastive expansions are extracted from a large corpus of queries submitted to the Blekko web search engine. The corpus contains 450 million distinct queries, along with their frequencies.

Candidate filtering. We then apply to the extracted expansions a set of unsupervised filters, which aim to exclude erroneous expansions. For example, we filter (**topic**, expansion) pairs with low semantic similarity/relatedness or incompatible corpus frequencies, as well as pairs where one concept is a sub-string of the other. For consistent expansions, we computed semantic similarity based on cosine similarity between word2vec representations. For contrastive expansions, we utilized WORT.

Candidate classification. Finally, we employ supervised classification to identify good expansions amongst the candidates. For each expansion type we train the following two complementary classifiers, which are combined by summing their output scores:

1. A logistic regression classifier, for which we developed a novel set of features. In addition to semantic similarity/relatedness discussed above, these features are based on Wikipedia metadata (e.g., number of shared categories and outlinks), WordNet relations, and sentiment and corpus statistics.

2. A neural model, which is trained by distant supervision [29]. For each instance *(**topic**, expansion)* $\Rightarrow L$ in the training data, where L is the label, we extract from LexisNexis a set of sentences that contain both concepts, and mask their occurrence in the sentences. The sentences collected for the whole training set are then used to train a neural model. Essentially, the network aims to determine whether a given sentence is a positive or a negative evidence for the existence of the target relation (consistent or contrastive expansion) between the masked concepts. When applying the classifier to a new pair, we collect up to 500 sentences for that pair, and average the classifier's predictions for each sentence. We use the same network architecture as was used for Sentence ranking in Section 5.1.1.

5.2 The Argument Knowledge-Base

The core of the Argument Knowledge-Base (AKB) is the definition of Classes of Principled (or reoccurring) Arguments (CoPAs), as described in [30]. We define various content types, described below, with each class containing a set of texts of each type.

For example, one of the classes is *Black Market*, and contains arguments such as: *Prohibition is counterproductive. Banning [TOPIC] increases demand for [it/them] by creating the 'forbidden fruit' effect. This results in precisely the opposite of what the ban intended to achieve,* which suggests that regulation, rather than banning, is the way to deal with the danger of a black market. By contrast, it also contains the argument: *Prohibition of illicit substances and services makes them less available and thus less harmful.* As can be seen from the first example, the texts may include placeholders which are filled in according to the specific debate in which they are applied. In addition, they may contain prosodic cues for the text-to-speech system, such as what words should be emphasized. Furthermore, the examples above demonstrate that classes contain texts reflecting the two opposing views pertaining to the underlying principled issue. Hence, it is not enough to match a motion to a class, one also needs to understand its stance toward it.

To do this, each CoPA is accompanied by a list of terms alongside their polarity w.r.t. the class. These terms help determine whether or not a class is applicable for a motion (see [30]), and, if it is, what stance to select. For example, for the Black Market class, terms which indicate a regulatory approach include Legalization and Regulation; terms which indicate a prohibitive approach include Prohibition and Black market; terms which are indicative of the class but not of a clear stance include Forbidden Fruit and Crime. In addition, the **action** is often directly indicative of the stance. Ultimately, both motion-class matching and stance classification are done by a funnel of classifiers.

Importantly, matching a motion to a CoPA is far from trivial. For example, one might think that the *Black Market* CoPA is applicable to any motion where the **action** is *ban* or *legalize*. However, considering motions such as *we should ban breast feeding in public* and *we should legalize polygamy*, shows that this task is often complex and nuanced. Indeed, to attain a good recognition of this class one of the main methods we used relied on leveraging direct world knowledge. For example, when presented with a novel motion, one of the cues the system attains for whether or not Black Market arguments are relevant is by searching our LexisNexis corpus for sentences which mention both the **topic** and the concept Black Market (see [30] for a description of some of the other features).

These are the types of contents a CoPA may include, alongside an example from the Black Market CoPA (those related to rebuttal are deferred to section 5.2.1):

1. **Claims**, e.g., *A licensed and regulated environment is the only way to protect individuals and minimize harm.*

2. **Extended claims**, e.g., *Regulation is needed to protect individuals from the dangers of a black market. Banning is not the solution, rather a licensed and regulated environment is the best way to protect those individuals and reduce the potential harm done to them.*

3. **Evidence**, e.g., *Although popular opinion believes that Prohibition failed, it succeeded in cutting overall alcohol consumption in half during the 1920s, and consumption remained below pre-Prohibition levels until the 1940s, suggesting that prohibition did socialize a significant proportion of the population in temperate habits.*

4. **Opening or Framing**, e.g., *In this debate we will be talking about responsibility. We all agree that as much as we would like to, we can't trust everyone to be the perfect citizen and some control is needed in those areas that call for it. Even if there are downsides to banning, it is sometimes the only viable alternative.*

5. **Conclusion referring to the opening**, e.g., *In my opening speech I talked about responsibility. And I hope that since that moment, I gave you enough evidence and reasons to agree that sometimes, like with [TOPIC], banning is the only viable option.*

6. **Proactive arguments**, e.g., *My opposition today will in all probability side with the strict camp of banning. I would like to ask a simple question, [OPPONENT NAME], do you really think that laws would stop [TOPIC] ?*

7. **Motion-independent response**, e.g., *The risk that banning will cause some unwanted temporary side effects, like the rise of a black market, may exist to some extent. But we*

should also bear in mind how effective banning is. Prohibiting something makes it less visible and available, and thus less harmful. Hence, the overall impact is still clearly positive.

8. **Quote**, e.g., *The more we bring these things into the light, you bring these hidden vices into the light, the less power they have over our society. You make it legal, it tends to go away. (Actor Thomas Jane)*

9. **Humorous similes**, e.g., *Banning [TOPIC] is like mending cracks in a wall. It may not be the perfect solution, but it is necessary.*

10. **Closing words**, e.g., *In conclusion, what opponents basically suggest is to give up and let society handle matters itself. Let people do what they need to do, even if it harms them and others. We disagree. We believe in the right of the government to protect its citizens and uphold the values it stands for.*

In addition, the AKB contains some content which is not part of the class system. This includes the following:

1. **Action-related Example**, which is used when the topic of the example is related to the topic of the motion. For example, when the **action** is "ban", the following text may be used if the **topic** is similar or related to smoking: *New research by the University of Glasgow suggests that smoking bans across the UK have reduced the uptake of smoking by teenagers by roughly a fifth.*.

2. **Keyword-triggered response**, which is used when the opponent mentions the keyword, and no other response can be generated. For example, if the opponent mentions "cost" this "last-resort" response might be *My opponent referred to the issue of cost in reference to [TOPIC]. Indeed, costs affect us all. However, we must not forget that costs are not the*

be-all and end-all of everything. Today we are highlighting other aspects of life, that are just as important.

3. **Framing based on action**. Analysis motions differ from policy motions in that they do not deliberate a specific policy w.r.t. the **topic**, but rather analyze its inherent merits and flaws. For such motions, text expressing this is added, based on the **view**. For example, for the **view** *is justified*, this text is *Importantly, the aim of this debate is to discuss the* ***justification*** *of [TOPIC], rather than specific policies that might follow from it being justified.*

5.2.1 The Argument Knowledge-Base for Rebuttal

The part of the AKB pertaining to rebuttal is based on a *Rebuttal Unit* composed of three parts:

1. What to search for in the opponent's speech (*lead*).

2. How to reference the opponent's argument (*reference text*).

3. How to respond to the opponent's argument (*rebuttal text*).

Given a *Rebuttal Unit*, the system works as follows: It searches the opponent's speech text for sentences that mention the *lead*. Roughly, it examines each sentence in the speech, and determines whether the sentence implies the lead. If any of them do, it suggests using a text along the following lines by way of rebuttal:

> My opponent claimed that [*reference text*]. However, [*rebuttal text*].

For example, a *Rebuttal Unit* may correspond to the Black Market argument. The *lead* might be the text "black market". The *reference text* might be "A black market will arise", and the *rebuttal text* might be "Even if there is some truth to that, it should not stop us from doing what is right". When the system deems such a *Rebuttal Unit* as relevant, it will try to determine whether the opponent mentioned the term “black market”. If so, the system will suggest embedding in the speech a paragraph similar to:

> My opponent claimed that a black market will arise. However, even if there is some truth to that, it should not stop us from doing what is right.

Leads defined in the Knowledge-Base are either short claims [31], short phrases or sentiment words. The same mechanism is also used for rebuttal based on iDebate, themes identified in the opponent's speech and mined arguments. In the latter two cases, the *lead* is not pre-defined, but depends on what the opponent said or on the results of the argument mining components. Section 5.3 describes the rebuttal system in detail.

5.3 The Rebuttal System

The goal of the rebuttal system is producing rebuttal arguments that respond to claims argued by the human opponent. The identified opponent claims and their rebuttal texts are passed to the Debate Construction component, which integrates them into a full speech that also includes other non-rebuttal content (see Section 5.4).

Processing starts with converting the opponent's speech audio into text using IBM's Watson's off-the-shelf Automatic Speech to Text (STT)[2]. The STT output is automatically cleaned, for example by removing the timings information and non-textual information such as hesitation marks. The obtained word sequence is automatically punctuated and split into sentences using a bi-directional LSTM [32]. This text is the input to several independent components, described below, aimed at understanding its contents.

A prerequisite to the development and evaluation of components that produce rebuttal arguments to opponent speeches is having suitable data, simulating their inputs in live debates, namely, a corpus of debate speeches, each including one speaker arguing for or against a given

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

motion. Such data was recorded at scale by a team of expert debaters. To produce a debate speech, a debater was presented with a motion and its description, and was instructed to record a four minute speech that supports that motion, with 10 minutes to prepare, but without checking any online materials. Given a speech supporting a motion, a fellow debater listened and recorded a speech rebutting it, and in particular – opposing that motion. Overall, more than 3,600 speeches were recorded, discussing more than 400 motions. Most of that data was released in [33], and a detailed account of its collection procedure is given in [34].

Several of the rebuttal components work by determining which short claims (or *leads*) from a predefined list were indeed mentioned by the opponent, and suggesting an appropriate rebuttal response to those identified as mentioned. These components are trained and evaluated on a collection of speech and claim pairs annotated for whether the claim is mentioned or not in the speech. Initially, such *speech-level* data was annotated by experts (e.g., in [35]), however relying on experts alone limited the amount of data which could be collected. Adjusting the task to the crowd was not trivial – annotating full speeches, and long lists of leads potentially mentioned in those speeches, is on a different level of complexity in comparison to annotations done for simpler NLU tasks, which most often involve shorter texts. For example, the data collected for the tasks in the GLUE benchmark [36] involves one sentence or paragraph in each labeled instance. In [37] we proposed a scheme for effectively performing this complex task with crowd annotators by applying various quality control measures that identify reliable annotators qualified to work *on our specific data*. We believe that the encountered challenges and pitfalls, as well as lessons learned, are relevant in general when collecting data at scale for complex NLU tasks.

A second type of data is aimed at understanding where in a speech a lead was mentioned. This *sentence-level* annotation includes pairs of lead and opponent speech sentence, annotated for whether the lead is mentioned or not in the sentence. Naturally, there are many such pairs since each speech is comprised of about 30 sentences and has tens of potentially mentioned leads, so it is unfeasible to comprehensively annotate all pairs, and sampling is required when selecting the annotation inputs. Since randomly sampling a lead and a sentence is likely to produce a pair in which the lead is not mentioned in the sentences [37], the annotation included the top-ranked leads predicted for each speech, from one of the methods described below, along with the sentences in which they were presumably identified. More than 6,000 annotated lead and sentence pairs are available as part of the dataset introduced in [31]. A comparison of the sentence-level and speech-level annotation schemes is included in [37].

Next, we provide a description of the rebuttal components that are included in the rebuttal module.

5.3.1 Key-Concepts Identification

The first component aims to identify the key concepts discussed by the opponent. It starts by running the Wikifier to extract mentions. Those mentions are scored by a logistic regression classifier, trained on data collected for this component: speeches and mentions automatically found in those speeches, annotated for whether the mention is a key concept in the speech. The features used by the classifier include, among others: (i) semantic similarities such as the similarity between the **topic** and the concept that a mention was mapped to, using WORT; (ii) count based features such as the count of the mention (or concept) in the speech, optionally normalized by the total number of mentions (or concepts) in the speech, or weighted by the inverse document frequency of the concept in Wikipedia; (iii) binary features such as whether the mention contains digits or punctuation marks. The output of this component, namely, the identified key concepts and their assigned scores, are used by downstream components – e.g., by features of the classifier responsible for detecting AKB leads in opponent speech sentences (see Section 5.3.4 below).

5.3.2 Claim Leads and Counter-Evidence Detection

As described in Section 5.1.1, an important capability of Project Debater is automatic mining of claims from a large text corpus, enabling the system to present high-quality content supporting its side.

For rebuttal, this capability is utilized as well: mined claims, identified to support the opponent's side are searched for in the opponent speech using trained classifiers (see [38] for details).

To rebut a claim identified as mentioned by the opponent, we try to match it with counter evidence (see Section 5.1.2). We run the Wikifier on the claim, as well as on the evidence, to extract Wikipedia concepts therein (excluding the **topic** and an additional small set of manually pre-defined concepts). Finally, we sort the evidence, by calculating the ratio of Wikipedia concepts in the claim that are also found in the evidence, requiring a minimal ratio of 30%. The output of this component is thus a mapping between predicted opponent claims and counter-evidence, sorted by their relevance, which can then be used as needed for rebuttal.

5.3.3 Sentiment Leads

Human debaters often use a simple phrasing to mention a positive or a negative effect of the topic (e.g., "*preschool can be harmful* for some children"). We aim to capture such statements by looking for *sentiment leads* – statements with a clear sentiment towards the topic – in opponent speeches.

For a given **topic** (e.g., "preschool"), we generate sentiment leads of the format *topic is/are sentiment*, where *sentiment* is a sentiment bearing word from a short pre-defined list. When the opponent supports the topic we look for positive sentiment words (*valuable*, *beneficial*, *helpful*, *preferable*, *safe*, *effective*), and when the opponent opposes the topic, we look for negative sentiment words (*dangerous*, *immoral*, *harmful*, *abusive*, *unjust*, *inefficient*, *problematic*, *arbitrary*, *expensive*). These lists were constructed manually by analyzing speeches of human debaters.

In an opponent's speech, we look for speech sub-sentences whose semantic similarity (determined by word2vec [27]) to one of our sentiment leads is above a threshold (tuned on development motions). When such a sub-sentence is found, we deduce that the opponent has mentioned the lead. For example, the sub-sentence "preschools are going to be very expensive" is semantically close to the lead "preschools are expensive". An evaluation of this method over about 600 speeches (from [33]) yielded a precision of 76%, and a speech coverage of about 12% (the percentage of speeches with at least one predicted sentiment lead).

For these general sentiment statements, identified in opponent speeches, we do not have a specific counter argument prepared in advance. Therefore, we use one out of 5 simple templates to generate a general response that mentions the opponent statement and shifts the discussion towards the arguments we prefer to present. Continuing our example, the response to the lead "preschools are expensive", when identified as mentioned by the opponent, could be ''*My opponent claimed that preschools are expensive. I believe my arguments suggested that the benefits outweigh the potential disadvantages*". This technique, which does not directly rebut the opponent argument, is called blind prioritization and is a simple method human debaters use to prioritize their own arguments when they want to move the debate to an area which is strategically beneficial for their case.

5.3.4 Claim Leads from the Argument Knowledge-Base

As described in Section 5.2.1, the AKB includes Rebuttal Units, each containing a lead to be searched for in the opponent's speech, along with the appropriate rebuttal text, to be used as a response when the lead has been detected in a speech. One type of such leads are phrased as short claims, such as *Adopting this proposal would lead to backlash.* This section describes the various components that are involved in their detection.

Prior. An important property of the AKB leads is that the same lead can (and does) appear across different motions and speeches. Specifically, given a training set, the a-priori probability that a lead will be mentioned in a speech can be computed. Then, given a test speech, leads are scored with their computed a-priori probability, without considering the text of the speech. Focusing on leads with a high prior yields high precision [31], at the cost of an undiversified output, since the same high-prior leads are reoccurring within that output.

Nearest-Neighbours. Each lead has an associated sentence-level annotated data, comprised of a set of sentences (from opponent speeches recorded for development motions) annotated for whether that lead is mentioned or not in each sentence. These data can be used to determine whether a lead is mentioned in a test speech, by examining the speech contents and identifying whether parts of it are semantically similar to annotated sentences in which the lead is annotated as mentioned.

Specifically, each test speech sentence is vector-represented by calculating the weighted average of the embeddings of its words, using IDF weights based on Wikipedia. The annotated sentences associated with the lead are represented similarly. Next, the similarities between the test speech sentences and the annotated sentences are computed using the cosine similarity. Focusing on one sentence from the test speech, its K-most similar annotated sentences are taken, and if their majority was annotated as mentioning the lead, then the test sentence is also considered as mentioning it. The similarity score between the test sentence and the lead is set to the similarity between the test sentence and its nearest annotated neighbour that mentions the lead. Finally, maximizing those similarity scores over all sentences in a test speech yields a lead-to-speech similarity score, which is used to rank leads detected in this manner.

Logistic Regression. The sentence-level data is also used to train a logistic regression classifier, aimed at predicting whether a lead is mentioned in a given sentence. Its features include various text similarity measures, which were also used for identifying mined claim leads (see [38] for details). These were augmented with features that utilize the additional information available in the AKB for each lead (information that does not exist for mined leads). For example, one such feature computed the percentage of concepts identified in an input sentence that are also present as concepts for the lead in the AKB.

Neural Model. A neural model is trained for the same task as the logistic regression classifier: determine whether or not a lead is mentioned in a sentence. The inputs to the classifier are a pair of lead and sentence, and the a-priory probability of the lead to appear in the speech. The lead and the sentence are embedded with two different bi-LSTMs. We add an attention layer to the concatenation of those two representation, and concatenated the prior. On top of that layer we add a fully connected layer and finally apply softmax to predict the output.

Keywords. A subset of the leads is annotated with a per-lead list of keywords. Using those lists, a sentence from an opponent speech which contains all the keywords listed for some lead is predicted as mentioning that lead. For example, the lead *Adopting this proposal would lead to backlash* may be searched for using the keyword *backlash* alone.

The precision obtained by each keyword list was estimated on development motions, by annotating the leads and the sentences that were predicted as mentioning them. Keyword lists with a precision above a desired threshold are activated for test motions. Others are shown to in-house annotators along with the results of their annotation (which is a list of sentences annotated for whether the lead is mentioned in them), and the annotators are asked to improve those lists, if possible. These, in turn, produce new sentences in which the leads are identified, which are then sent again to precision estimation, and are either activated in test, sent again to manual revision, or disabled. The entire process may be repeated for several iterations. The high involvement of humans-in-the-loop in this method made its use at scale too costly, so it was only applied to few leads. For those, it provided a simple, high-precision solution.

Estimating Precision. The precision of the predictions of each component can be estimated for each lead by running the component on the development motions in a leave-one-motion-out cross validation mode. In each fold, training is on all motions except one, and precision is evaluated on the left-out motion. Averaging over all folds yields the per-component estimated precision of each lead. For each component, a threshold is defined on the estimated precision of its leads, so leads with an estimated precision below this threshold are never predicted by this component, regardless of the opponent's speech contents. This is similar to the prior-based component, yet different in that the a-priori probability for success is computed on the predictions made by each component on the development set, rather than on the speech-level labeled data. For example, low-prior leads, which are identified with a high precision by one of the components, are included in the output so long as that precision is above the predefined threshold for the corresponding component.

Aggregation Lastly, predictions from different components are aggregated together. The final match score of each lead with respect to the input speech is set to its maximum match score from all available components (prior, nearest neighbours, logistic regression, neural model, and keywords). Next, since different leads in the AKB may have the same pre-written response, to avoid including the same response twice in one speech, only the top-scoring lead of each available response is selected. Then, the leads are sorted by their similarity score to the opponent speech, and each lead is added to the output, unless a similar lead is already present in the output. Lead similarity is calculated by finding for each word (excluding stopwords) in one text its most similar word in the other text (where pairwise word similarity is defined by the cosine similarity between their corresponding word2vec embeddings), and averaging those similarities.

Attaching Detected Evidence. The pre-written AKB response could be better connected to the motion by adding to it an evidence relevant for the topic, which is suitable to the response (for evidence detection see Section 5.1.2). For example, in the context of the motion *We should further exploit hydroelectric dams*, and the AKB-based response arguing against it – *hydroelectric dams might be good for the environment, but it is not worth the investment*, the response is assume to be better connected to the motion when followed by the mined expert evidence – *Experts say that the government should deemphasize investment in large hydro and increase investment in sustainable, decentralized renewable sources*. Such an evidence, immediately following an AKB response, is referred to as "attached" to the response.

The first step in identifying an evidence which can be attached to a given AKB response, is detecting the Wikipedia concepts mentioned in the evidence using the Wikifier. The AKB response is similarly represented by a bag of concepts manually extracted from its text, that are saved in the AKB. The similarity between each pair of evidence concept and response concept is computed using WORT. These similarities are aggregated, by finding for each evidence concept its similarity score to the most similar AKB response concept, and doing so for the AKB response concepts as well. Averaging those similarities gives the similarity of an AKB response and an evidence text.

This component was evaluated on pairs of pre-written AKB response and evidence, in the context of a specific motion, annotated for whether their joint use is preferable to using the AKB response by itself. As with the lead identification components, it is possible to estimate the probability of successfully attaching an evidence to each AKB response, using the development data. Some AKB responses had no evidence annotated as a valuable addition when attached to them. Hence, to these responses, the system avoided attaching evidence at test time.

5.3.5 iDebate-Based Rebuttal

Another potential rebuttal resource is iDebate[3], a manually curated database of motions, containing a list of arguments for and against each motion. Each argument is coupled with a counter-argument, which may be used as a response once the argument has been established as mentioned by the opponent. Naturally, this resource can only be used for motions included in it, hence has inherently limited coverage. Yet, it is advantageous due to the high-quality of its counter-arguments. Therefore, a response from this resource was prioritized over other forms of rebuttal, in the rare cases in which it was available. However, we limited Project Debater to use at most one response from this resource in a full debate.

The component starts by checking whether the motion being discussed is present in the database. In a pre-processing step, an in-house annotator extracted the topic of each motion in iDebate, and mapped it to a Wikipedia concept, when such a concept exists. The annotator also identified whether the phrasing of the motion supports or contests the discussed topic. Then the **topic** is compared to the list of concepts associated with iDebate motions. Only exact concept matches are considered, since a matching mistake, resulting in the use of arguments from a semantically similar yet irrelevant motion, was deemed unacceptable. That further limited the set of motions which could potentially benefit form this resource for rebuttal, but ensured the precision of this first step. In our development motions set, we had a matching to an iDebate motion for $\sim$10% of the motions.

After a relevant iDebate motion has been identified, its stance towards the topic is compared to the stance of the input motion towards that same topic, and its arguments supporting the opponent's side are taken. For example, if the input motion *We should ban cannabis* was matched to the iDebate motion *This House believes that cannabis should be legalised*, then arguments supporting the iDebate motion contest the input motion and thus potentially could

[3] `http://idebate.org/`

be mentioned by the human opponent. Each potentially mentioned iDebate argument may be comprised of several sentences and is briefly summarized in a title, which is similar to a claim in its nature. Continuing our example, *If cannabis was legalized, it could be regulated* is such a title supporting the aforementioned topic. Those titles are therefore considered as leads, and their similarity to a given opponent speech is computed in an unsupervised manner, based on word-embeddings (see [35] for details). A threshold is applied to that similarity score to obtain the desired precision.

When an iDebate argument title is identified as mentioned by the opponent, its response is produced by taking the first sentence of its counter argument, when the length of that sentence is within a predefined range of 15-70 tokens. In cases where the first sentence was too short, the second sentence was also used, if their combined length was under 70 tokens.

5.3.6 Detecting Opponent's Main Themes

Professional debaters often explicitly declare their main themes to make it easier for the audience to follow them (Project Debater uses themes as well, around which speech paragraphs are constructed, as discussed later in Section 5.4.4).

We developed a component to detect the main themes in the opponent's speech. As speeches annotated with their main themes are not available, a supervised machine learning approach was infeasible. Instead, we extract themes with manually written rules following analysis of speeches from development motions. The rules take advantage of the structure that opponents often introduce into their speeches to make them easily understandable by their audience. These rules look for sentences which include indicators for a statement of a theme (e.g., "*My first argument is simply the notion of the duties of the state to every man and woman and I will claim that...*"). Next, other rules are applied, on the text which follows the indicator, to extract and rephrase the potentially mentioned theme-phrase. A valid theme-phrase is a concise noun-phrase (i.e., in the above example, *the duties of the state* as opposed to *simply the notion of the duties of the state*) or a nominal verb (e.g., *book reading* instead of *read a book*).

We took the approach of manually written rules[4] to meet the requirement of very high precision (prioritized over recall), since in a live debate scenario it is better to leave a point unopposed than oppose a point the opponent never said. Indeed, an analysis of almost 600 opponent speeches (from [33]) shows that we are able to identify a main theme in about 20% of the speeches, with a precision of 93%.

When an opponent theme is found, it is used during speech construction to identify a cluster with a similar theme, from within the available argument clusters constructed for Project Debater's side (see Sections 5.4.3 and 5.4.4). If such a cluster is found, it is selected to Project Debater's response speech, and is preceded by an opening sentence quoting the detected opponent theme, as in "*One issue my opponent mentioned was the duties of the state. My arguments suggested there are more issues to consider*". Then, Project Debater's arguments from that similarly-themed cluster are presented. If no cluster has a theme which is similar enough to the identified opponent theme, we use the identified theme in the rebuttal paragraph in a similar manner to our use of sentiment leads (see Section 5.3.3).

This rebuttal type is different from those presented above, as it does not search for a predefined set of leads within the opponent's speech. Its output themes can come from an open set of possible phrases that are unknown in advance, thus enabling rebuttal with a stronger on-the-fly flavor.

5.4 Debate Construction

Next, we outline how all the pieces described in the preceding sections are combined to form a coherent speech. Specifically, the building blocks at our disposal are ranked lists of claims

[4] In a followup work [39] we used the GrASP algorithm [40] to automatically acquire such rules.

and evidence retrieved from the LexisNexis corpus, which are predicted to support the government side; the Rebuttal Units; the various CoPA-based texts which include placeholders for motion-dependent terms, such as the **topic**; and examples from other fields. In addition, the system includes boilerplate texts (which, like CoPA-texts, may also include placeholders), and a definition for the **topic** extracted from Wikipedia.

The debate construction pipeline is comprised of the following stages:

5.4.1 Element Filtering

The LexisNexis claim and evidence lists are ranked by their respective neural model scores. For the debate construction, we would like to make use of only a couple of dozen top ranked candidates.

Therefore, we need a threshold to filter elements by their score. However, setting a general threshold is problematic for two reasons. First, scores are not necessarily comparable across motions. Second, the value of the scores, given by a neural model (which account for the entire evidence score and half of the claim score), do not have a clear probabilistic interpretation.

Therefore, for a given ranked list of elements, we would like to find several thresholds denoted threshold$_k$, $k \in \{0.6, 0.7, .0.8, 0.9\}$, such that, the set of all elements above it would have a precision of at least $k\%$ (e.g., at least 80% of the elements with score above threshold$_{80}$ are valid). We developed two algorithms for this purpose. The claim detection component finds these thresholds for every development motion in the leave-one-motion-out cross validation mode. For each left-out motion, all candidate claim elements of the other motions are sorted by their scores in a descending order. Next, for each threshold k we find the minimal score for which the precision of all candidate claim elements above it is at least $k\%$. We do this for every motion in the development set, and as the final threshold$_k$ take the median of all development motions threshold$_k$. The evidence detection component used a different algorithm. Specifically, we trained a linear regression for each target precision k. Every motion is a training example with a single feature – the average score of the top 30 candidate evidence elements – and the label is the minimal threshold score which guaranties a set of elements with precision not lower than $k\%$.[5]

Once thresholds are learned, ranked element lists for a motion can be divided into precision bins. These bins are comparable across motions and have a clear probabilistic interpretation. We manually select the minimal bin for claims and evidence, by analyzing, on a validation set, the quantity and quality of the elements that reach the speech from that precision bin.

Finally, elements are also filtered for low confidence of the stance detection task, for being too short or too long, and for inappropriate content or style (e.g., the text contains a word from a list of inappropriate words/phrases, or too many non-English symbols).

5.4.2 Repetition Removal

Often times, claims and evidence retrieved from different articles in LexisNexis make the same point, raising the need for removing this redundancy in advance. To this end, we employ a repetition removal mechanism, that clusters similar arguments together and keeps a single representative within each cluster for downstream tasks.

This mechanism is a variant of agglomerative clustering, and operates as follows: first, each argument is represented by the mean of the word2vec vectors of its tokens, weighted by their IDF in Wikipedia. Initially, each argument is mapped to a singleton cluster. At each step, the algorithm attempts to merge the two most similar clusters, using cosine similarity as a similarity metric, as long as all arguments in the merged set are sufficiently similar. Finally, a single argument is selected per cluster as the cluster representative, while the rest of the arguments are discarded.

[5] Various features and feature combinations were tested, but the average score of top 30 elements was found to be the best predictor.

5.4.3 Clustering

The remaining representative arguments are clustered into thematic clusters, enabling to present arguments pertaining to the same theme in a single cohesive paragraph. For example, in a motion where there are arguments that touch upon economic subjects and health concerns, the clustering algorithm should find a partition that separates between these two sets of arguments for the purpose of presenting each set separately. Clustering is done using the iClust algorithm [41]; Wikifier is used to represent each argument by its respective Wikipedia concepts; and WORT is used as a similarity metric.

5.4.4 Theme Extraction

The output of the Clustering step is a set of clusters where each presumably discusses a specific subject or *theme*. The goal of *Theme extraction* is to identify the main *theme* of each cluster, alongside an optional theme-claim (*t-claim*). When generating the speech, we use these entities to prepare an introduction to argumentative paragraphs (see Table 3).

The *theme* is a Wikipedia title, typically a noun or a noun phrase (e.g., *Economy* or *Health*), which aims to describe the high-level subject that the arguments of the cluster address. The theme is extracted using a combination of global and local measures. As a preliminary step, the Wikifier is used to identify all Wikipedia concepts in the arguments of all clusters. Then, we use the hyper-geometric test to estimate the enrichment of each title in each cluster, preserving only titles that are statistically enriched ($p < 0.05$) in a particular cluster. In addition, in each cluster we count the number of arguments that mention the title, or a similar Wikipedia title, where similarity is determined by a pre-defined threshold score of WORT. By combining these measures, we ensure that the selected theme of a cluster is both frequent in that cluster, and relatively unique to it.

The *t-claim* is an expansion of the theme, which aims at exemplifying how the **topic** and the theme are related. For example, for the motion *We should subsidize higher education*, a t-claim for a cluster with the theme *Economy* could be: *Higher education boosts the economy*. T-claims are sub-sentential fragments that are extracted from the cluster's arguments, based on syntactic rules. If multiple t-claims are found for a given cluster (theme), the shortest one is used in the speech.

5.4.5 Rephrasing

Often, simply concatenating a list of arguments drawn from multiple contexts, does not give rise to a fluent text. Thus, some rephrasing is required to modify the arguments and produce a more compelling output. We distinguish between (a) rephrasing that works at the argument level, such as cleansing of connectives and anaphora resolution; and (b) rephrasing that works at the paragraph level, such as concatenation of a list of short arguments into a single, longer argument, or replacement of a noun phrase that is repeated in a list of arguments by a pronoun that refers to the first mention.

Next, we describe two argument-level rephrasers that exemplify the challenges in maintaining coherence, clarity, and fluency in a speech relying on mined arguments.

The first rephraser adds a description of people mentioned in the mined arguments. This is required in order to give the listener some details about the person being quoted or referred to, aiming to make the argument more credible and self-contained. Table 1 illustrates this rephraser on an argument for the motion *We should subsidize higher education*. The text added by the rephraser (marked in boldface) is extracted from the full text of the document within which the argument originally appeared.

Before	**After**

Bryan Caplan argues that higher education “is a big waste of time and money” and that “students spend thousands of hours studying subjects irrelevant to the modern labor market.”	Bryan Caplan**, an economics professor at George Mason University** argues that higher education “is a big waste of time and money” and that “students spend thousands of hours studying subjects irrelevant to the modern labor market.”

Table 1: Adding person description

The second rephraser deals with the transition from written language to spoken language. In many cases, arguments that are well phrased in written language would sound unclear when spoken aloud. One such example involves the syntactic construction of a quote followed by a *said by* phrase. This structure, common in news articles, is confusing when used in spoken language, where it would be more natural to first mention the person being quoted. Table 2 demonstrates the impact of this rephraser on an argument for the motion “We should limit the use of birth control”. The rephraser moves the *said by* phrase to the beginning of the text.

Before	**After**
“Scientists have abundant evidence that birth control has significant health benefits for women and their families, is documented to significantly reduce health costs, and is the most commonly taken drug in America by young and middle-aged women,” **U.S. Department of Health and Human Services Secretary Kathleen Sebelius said in a 2012 statement**.	**U.S. Department of Health and Human Services Secretary Kathleen Sebelius said in a 2012 statement that** “Scientists have abundant evidence that birth control has significant health benefits for women and their families, is documented to significantly reduce health costs, and is the most commonly taken drug in America by young and middle-aged women”.

Table 2: Advancing a *said by* phrase

Overall, the system makes use of 62 different rephrasers, some of which perform simple string substitutions, while others perform more complex operations, as in the two examples above.

5.4.6 Speech Generation

A speech is composed of paragraphs, which, in turn, are constructed using rules which depend on the type of the paragraph. These rules describe which of the underlying building blocks (e.g., claims, evidence, CoPA-texts) are used in each paragraph type, and how they should be combined. A speech is thus based on a list of typed paragraphs, and some rules on when to include and how to prioritize these paragraphs.

An important paragraph type is the one that includes clusters of minded arguments. Before generating the opening speech, the system ranks the clusters and selects the top ones. Ranking is based on many parameters, e.g., whether the cluster has a *t-claim*, the number of arguments in the cluster, and the average similarity between all argument pairs. The top scored clusters are then assigned to the opening and rebuttal speeches (up to 3 clusters per speech). During the speech generation, the system generates an *argumentative paragraph* for each assigned cluster.

Often there are many arguments in each cluster and one needs to choose which to use, how to order them within a paragraph and how to order the paragraphs themselves. The argumentative paragraphs are constructed by an iterative process that scores each argument in the cluster. The score takes into account the confidence in the argument stance, the confidence of the respective claim/evidence detection component, and its similarity to the theme of the cluster (based on Wikifier and WORT). The top scoring arguments are then selected into the cluster's paragraph.

Speeches are generated using pre-defined templates for each round of the debate – an opening speech, a rebuttal speech, and a summary speech – where each template serves as a skeleton that determines the respective speech structure.

Table 3 describes the potential paragraphs appearing in the system's opening speech, and the conditions under which they may appear. In addition to these conditions, paragraphs may be trimmed of content or removed altogether so as not to exceed the 4 minutes allotted to this speech.

Order	Paragraph	Example	Condition
1	Greeting	Greetings and thanks for the opportunity to participate in this debate. Based on my analysis, we should not abandon the Paris Agreement.	
2	Definition	Let me start with a few words of background. The Paris Agreement is an agreement within the United Nations Framework Convention on Climate Change dealing with greenhouse gases emissions mitigation, adaptation and finance starting in the year 2020.	**topic** is infrequent in corpus
3	CoPA-based opening	A big issue in today's discussion is the environment. Everyone agrees that clean air, water and energy sources are preferable to a polluted earth, and that it is worth investing in order to achieve and maximize these goals.	CoPA matched to motion
4	CoPA-based arguments	In harming the environment, the citizens of the world ultimately harm themselves. When they cause damage to nature, they endanger the ability of the human race to sustain itself in the future. In 2016, the renowned astrophysicist Stephen Hawking said in an Oxford University lecture that mounting environmental challenges and the depletion of natural resources means that humanity has at most one thousand years left on Earth.	CoPA matched to motion
5	Arguments intro - themes and t-claims	There are several issues I would like to address. They explain why we should not abandon the Paris Agreement. I will start by explaining why the Paris Climate Accord represents the best way to address climate change [...].	

6	Argument paragraph (may appear multiple times)	Starting with emissions. The Paris agreement introduces more robust greenhouse gas accounting. Surveys carried out by Yale University after the recent election found that 69 percent of registered voters support our participation in the Paris agreement and 66 percent support reducing greenhouse-gas emissions, regardless of what other countries do. Moreover, 47 percent of Trump voters agree.	
7	Arguments summary - themes and t-claims	Let me briefly summarize my introduction speech. I argued that the Paris agreement is a global framework for protecting prosperity [...].	
8	Pre-closing	Thus, my understanding is that we should not abandon the Paris Agreement.	
9	CoPA-based proactive argument	My co-debater today will likely say that immediate human interests come first. I would like them to supply evidence showing how ruining the environment would benefit anyone in the long run.	CoPA matched to motion; proactive argument exists in CoPA
10	CoPA-based closing	In conclusion, I will repeat the point I started with: People have a duty to be the custodians of nature and to minimize the damage they inflict on the ecosystem.	No proactive argument
11	Closing	That concludes my speech. Thanks for listening.	

Table 3: Paragraphs of Opening Speech, exemplified for the motion *We should not abandon the Paris Agreement*.

The rebuttal speech is constructed in a similar manner. The main difference is that it includes a rebuttal paragraph, responding to the opponent's argument. For example, continuing our motion example above, such a paragraph might look like:

> First, I will address some of the points made by John Smith. I think that one of the claims made by John was that the Paris agreement has failed. The Paris agreement certainly has flaws in its implementation, but those flaws are not essential to its mission or construction. Making changes in organizational structure and leadership are a more efficient way to fulfill the same goal.

In addition, as an attempt of adding a comic relief, if the opponent seems to be speaking quickly - as computed by the total number of words in the speech divided by the speech's duration - the system will precede the rebuttal with a corresponding paragraph, e.g., :

> A comment, if I may. I couldn't fail to notice that you speak very fast. 217 words per minute, to be precise, which is quite above average. Please slow down, there is no need to hurry.

Another difference from the opening speech is in the CoPA-based content. It may include arguments from very general *special CoPAs* that are devised especially for this purpose, as well as a colorful metaphor related to a standard CoPA. Continuing the same example, these could be:

> Another point I would like to raise is the special impact this policy would have on developing countries. Take a moment to think about what not abandoning the Paris Agreement would mean in the context of the developing world, and how such countries in particular have a lot to gain from supporting this motion. (From the special CoPA *Developing countries*)

> Harming the environment is like shooting holes in your own boat. You will end up drowning. (From the CoPA *Environmentalism*)

The closing speech includes a rebuttal paragraph similar to that in the rebuttal speech, and a summary of the arguments made in the previous speeches. In addition, if the system detects a theme on which the two sides clashed, it will mention this before the argument summary (see also Section 9). For example, if the system identifies the *theme* Economy for the aforementioned motion, this paragraph might read:

> In this debate, we argued predominantly about the economy. Whereas John Smith brought forward his view that the Paris Agreement will destroy the economy, I had a different approach, saying among other things that the Pairs Agreement makes good economic sense.

Finally, the CoPA-based text in this speech includes a relevant quote and some CoPA-oriented concluding words. For example, these might be:

> Mankind has inflicted extensive damage on the environment. Yet, the fight is not over. We could save natural resources, flora and fauna if we put all of our efforts into it. Let's not give up, or we will also be remembered eventually for our brief existence on this planet.
> In the words of Mahatma Gandhi: "this planet can provide for human need, but not for human greed."

6 Baseline Systems

To compare the evaluation of the opening speech generated by the system to what can be attained by contemporary methods, we evaluate in the same manner several baselines. Our main points of comparisons are baselines which, like Project Debater, are fully automatic. These include two baselines based on GPT-2 [42] - one based on the generation of individual arguments and one on the generation of a full speech; a baseline composed of arguments extracted using the state-of-the-art argument mining algorithms of ArgumenText [43]; and a baseline generated via extractive query-based multi-document summarization using SUMMIT [44]. In addition, we examine two baselines which include manual human annotation - in one humans write arguments for the motion (collected in [18]), and in the other they review candidate arguments extracted by the system [15], and those which are confirmed by a majority of annotators are considered for the speech. Finally, we also compare the system to human performance for this task by evaluating transcripts of opening speeches recorded by 8 different expert debaters (two speeches per motion).

6.1 Baseline Speeches Based on Collections of Arguments

Four of the examined baselines are derived by collecting high-quality arguments pertaining to the motion, ordering them in a coherent manner and evaluating the resultant text in the same manner as Project Debater's opening speech. In all cases, arguments are ordered using the recent algorithm of [45]. This algorithm requires fine-tuning on a relevant dataset, for which we used transcripts of opening speeches from [33], excluding those motions on which these baselines were evaluated. Specifically, 198 motions were used for training, and 92 for validation. For each motion we selected one speech transcript at random. In each speech, we filtered out sentences shorter than 8 words or longer than 40 words. If the number of remaining sentences was larger than 10, we sampled 10 and deleted the rest. From the remaining sentences we generated all pair-wise comparisons yielding the training and validation sets. When presenting pairs of sentences to the classifier during training/inference we randomly set the order of sentences in each pair to prevent a situation when the classifier always sees the earlier sentence either as first or as second in each pair. The accuracy attained on the validation set was 0.65.

The four baselines differ in the arguments they use:

1. **Labeled Arguments Collected from the Crowd (Arg-Human1)**: 23 of the motions

analyzed here have a matching motion in [18], where arguments were solicited from crowd workers and annotated for quality by them. Only arguments supporting the motion are considered. Each argument is assigned a weighted average of these scores - defined as its *WA-score* in [18] - and sorted accordingly. Arguments with significant token overlap to higher-scoring arguments (Jaccard score > 0.8) are discarded. After this filtering, top scoring arguments are accumulated until a threshold of 600 tokens has been exceeded.

2. **Labeled Arguments from Newspaper Archive (Arg-Human2)**: For most of the motions we have collected, over the course of the project, we had human-curated arguments which were extracted from the LexisNexis corpus (as described in [15]). Arguments are typically reviewed by 15 crowd annotators, who provide a binary label for the appropriateness of the argument. Each argument is assigned a score equal to the fraction of annotators who labeled it as appropriate to the respective motion. Arguments which contest the motion are discarded. The remainder are sorted by this score, and then filtered and collected as above.

3. **Arguments Generated by GPT-2 (Arg-GPT2)**: GPT-2 [42] is a powerful transformer-based language model, trained to predict the next word in 40GB of Internet text. GPT-2 has exhibited remarkable quality of lengthy generated text, conditioned on some input [42]. Specifically, it has been shown effective when fine-tuned on down-stream text generation tasks (e.g, [46]). Here, we follow suit and train GPT-2 to generate synthetic arguments for a given motion, as recently described in [47]. To this end, we fine-tune the large GPT-2 model (774M parameters) for 1000 training steps on the dataset from [18], all supporting the motion. We consider arguments with a *WA-score* of at least 0.8, to maintain a high-quality training set. Initially, each argument is prompted with its respective motion. To provide additional context, we prepend to the prompt the first sentence of the

respective Wikipedia page. We concatenate prompts and arguments, separated by a delimiter. For example: *Influenza vaccines, also known as flu shots or flu jabs, are vaccines that protect against infection by Influenza viruses. Flu vaccination should be mandatory SEP The flu vaccine reduces the incidence of illness in the community.* For generation, we prompt the model in the same manner and generate several arguments, each limited by a length of 50 byte-pair encoding (bpe) tokens [48, 49], as initial experimentation has shown that very few generated arguments are longer. We generate arguments with top-k truncation (k=40) and a temperature of 0.7, maintaining a balance between coherence and diversity. Arguments are then filtered for redundancy and collected as above.

4. **Arguments extracted by ArgumenText (Arg-Search)**: The ArgumenText project [50, 43] provides APIs for multiple Argument Mining tasks, among them argument search within the common crawl dataset (from 2016). We have used this API to retrieve arguments for a query composed of a concatenation of the **action** and **topic** of a motion, using the default setting, except for the number of documents to retrieve, which was increased from the default 20 to 100, to better attain the required number of arguments. Arguments were divided to ”pro” and ”contra”, and the more appropriate list was selected manually. They were then sorted by their confidence score, and filtered and collected as above.

6.2 Full Speech Generated by GPT-2 (Speech-GPT2)

As described above, we considered GPT-2 for the generation of arguments which are building blocks for the construction of a full speech. However, GPT-2 is most impressive for its ability to generate synthetic text of arbitrary length, and as the examples in [42] show, it can adapt quickly to the style and content of the text it is conditioned on, even without any fine-tuning. Our initial experimentation, however, showed that a debate speech is far too restrictive and nuanced for GPT-2 to successfully imitate with zero-shot learning. Therefore, we fine-tuned the large GPT-

2 model on a dataset of recorded speeches [33]. Training data includes first round speeches from both the government and the opposition sides, prompted on the motion. For the latter, we flip the polarity of the prompt with simple rules (e.g., *We should disband the United Nations Security Council* is converted to *We should not disband the United Nations Security Council*), so that speeches are always in support of the prompt. We considered 240 motions which are not part of Pipeline-set-1 and Pipeline-set-2.

For the purpose of evaluation, we considered two fine-tuned models, which differ in their additional prompt context and generation sampling method. For additional context, we considered two alternatives: (i) prepending the first sentence from the respective Wikipedia page, as above; (ii) prepending the first paragraph therein, as presumably larger context could be beneficial for modeling long speeches. After concatenating each prompt with its respective speech, we filter samples with more than 1024 bpe tokens (GPT-2's limit), ending up with $2k$ and $1.5k$ training samples for (i) and (ii), respectively. We fine-tuned GPT-2 on (i) and (ii) for $2k$ steps each, obtaining two distinct fine-tuned models.

As generation sampling methods, we considered top-k sampling as well as nucleus sampling [51], which has been shown to better demonstrate a quality of human text in the context of story generation. After initial experimentation, we set nucleus sampling (p=0.9) with temperature=0.7 for model (i) and top-k sampling (k=40) with temperature=0.7 for model (ii). We then split the motions in the evaluation set between models (i) and (ii), such that each model generated speeches for half of the motions. The results that we present for Speech-GPT2 are based on generated speeches from both models combined[6].

[6] When analyzing the results, there was no significant advantage to either model.

6.3 Summarization (Summit)

The summarization baseline employed is based on SUMMIT [44], an extractive query-based multi-document summarization system. As this system requires documents to summarize, we query the LexisNexis corpus for documents with sentences containing the **action** and **topic** of the motion, while utilizing a lexicon of expansion for the former. We aim to obtain 1000 documents this way - documents beyond this number are discarded; if this number is not reached more documents are gathered to meet this threshold by looking for the word ”that” followed by the **topic** (as in [8, 6]).

SUMMIT then filters the sentences in these documents by keeping only a sentence that:

1. Contains the **action** (or one of its expansion terms) and the **topic**; or the word ”that” and the **topic**.
2. Is of length between 15 and 50 tokens.
3. Starts with a capital letter and ends with a full stop.
4. Does not contain a newline symbol.

The SUMMIT algorithm then selects a subset of these sentences based on the query string obtained by concatenating the **action** and the **topic**, and orders the selected sentences to produce a coherent summary.

7 Evaluation and Results

As described in the paper, we employ three types of evaluations: *Comparison to Baseline Systems* which can only be done for the opening speech; *Evaluation of the Final System* which evaluates the system as a debating system; and *Progress Over Time* which evaluates the system over the course of its development. Below we describe the statistical analyses employed followed by details on each evaluation method.

7.1 Statistical Analysis Details

In this supplementary and in the main paper we present several evaluation graphs, in which we present the scores of various systems (Figure 1 in the Supplementary and Figure 3 of the main paper), the score of Project Debater over two different sets of motions (Figures 2 and 5 in the Supplementary), and the score of Project Debater over several time points (Figures 3 and 4 in the Supplementary). For simplicity of presentation, we denote each bar in each of these figures as a *setting*. Thus, for example, 'Human Expert' is a setting in Figure 1, 'Eval-1' is a setting in Figures 2 and 5, and '2018' is a setting in Figures 3 and 4. Below we first describe how we calculate the empirical mean score for each setting, then how we calculate the error bars for these means, and finally how we calculate the significance when comparing the empirical mean scores of two settings.

Empirical Mean Score of a Setting

For each setting we consider N motions, and for each motion we present a specific question to n_A annotators, arriving at n_A scores per motion. The empirical mean score of a setting is calculated by first averaging the n_A scores of each motion, and then averaging over the N motions. We denote this empirical mean score of setting A by S_A.

Error Bars

The error bars around S_A represent the 95% Confidence Interval (CI) of S_A, based on bootstrapping. The bootstrapping is performed as follows: For each boot j, $j \in \{1, .., 1000\}$, we create a resampled set of labels per motion, by sampling with replacement n_A scores from the original n_A scores obtained for that motion. Thus, for each boot j, we have N sets of n_A labels. We then calculate the mean score of the j-th boot, S_A^j, analogously to the calculation of the empirical mean score. Based on the distribution of the obtained 1000 S_A^j values, we calculate the 95% CI of the empirical mean score. We note that, as expected, the mean of S_A^j is nearly identical to the empirical mean score, S_A.

Comparison of Two Settings

Given a pair of settings, e.g., Human Expert vs. Project Debater, denoted A and B, where A's empirical mean score is higher than B's, we consider the two sample problem, and ask whether the distribution of scores obtained in setting A is significantly higher than that obtained in setting B.

The null hypothesis is that the annotation scores in both settings originate from the same distribution. In order to create the null hypothesis distribution we employ the permutation method of [52] as follows: For each iteration j, $j \in \{1, .., 1000\}$, the n_A labels of setting A and n_B labels of setting B are pooled *per motion*, and divided randomly into two groups of size n_A and n_B (sampling without replacement). Thus for each iteration j, we have N sets of n_A and N sets of n_B permuted labels[7]. We then calculate S_A^j and S_B^j as above. The null hypothesis distribution is then defined as the distribution of the 1000 values of $D^j = S_A^j - S_B^j$, whose mean is ≈ 0.

We calculate the one-sided p-value of the significance test as the proportion of D^j values, such that D^j is greater than or equal to the difference between the empirical mean scores of the settings, i.e., $S_A - S_B$. For all significance analyses, we consider $p < 0.05$ as significant.

For the comparisons to the independent set of 36 motions (Figures 2 and 5), the null hypothesis is that there is no effect to the choice of motions. Here, we want to check whether the distribution of scores over one motion set is significantly different from the distribution of

[7] When the two settings do not have the same motions, e.g., the Human Expert setting has 77 of the 78 motions evaluated, we consider only the common motions. In this case, N denotes the number of common motions.

scores over a disjoint motion set. Therefore, we employ the permutation test, by pooling the motions, with their respective individual scores, and generating 1000 random partitions. The rest of the calculation is analogous to the one described above.

7.2 Comparison to Baseline Systems

7.2.1 Annotation Task

The opening speeches produced for each motion by Project Debater, various baselines, and human experts were presented using the Appen platform[8], in a random order, to crowd annotators selected based on exhibiting good performance on previous tasks. The annotators received the following instructions: *Imagine the following scenario. You are in the audience of a competitive debate between two opposing speakers on the specified topic. The first speaker delivers the opening speech, aiming to persuade the audience to support the topic. Please carefully read the transcript of this opening speech provided below. For each of the following statements, please indicate to what extent you agree or disagree with the statements.* They were then presented with the following statements:

1. This speech is a good opening speech for supporting the topic.
2. Most arguments in this speech support the topic.

For each, the annotators had to choose between 5 answers: ‘Strongly agree’; ‘Agree’; ‘Neither agree nor disagree’; ‘Disagree’; ‘Strongly disagree’. For further analyses, theses responses were mapped to a scale of 1 to 5 where 1 corresponds to ‘Strongly disagree’ and 5 to ‘Strongly agree’.

The guidelines reflect the scenario-based nature of the task, aiming to help annotators focus on the perspective of a debate audience. Further, the guidelines and questions do not reveal the

[8] `https://appen.com/`

possible origins of the speeches, so the annotators are in principle blind to whether the speech is produced by an automatic system or a human.

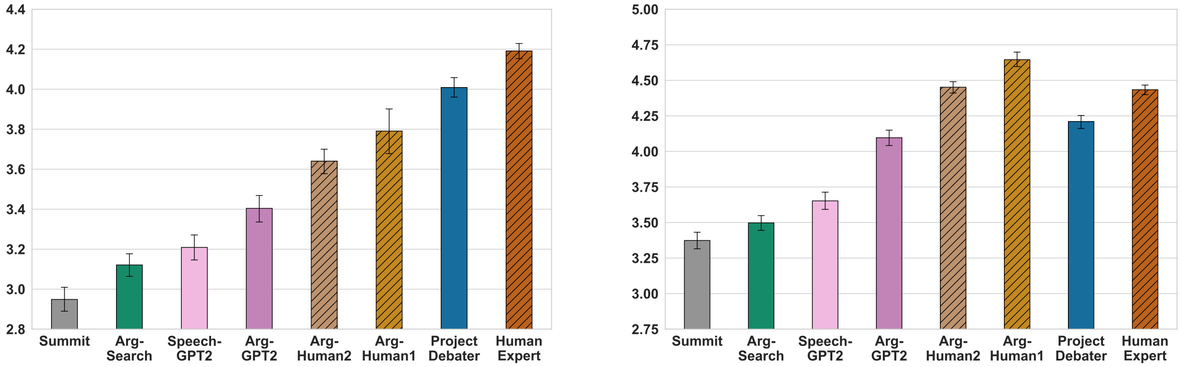
Each opening speech was annotated by 15 annotators. Overall, 84 annotators participated in the task, with 90% of the speeches annotated by a total of 49 annotators. The different motions and systems were mixed, and distributed to annotators by the Appen platform, ensuring that each annotator annotates at most 37% of the speeches.

7.2.2 Annotation Reliability

The inter annotator agreement for answering the first statement is $\kappa = 0.24$ as measured by quadratic weighted Cohen's Kappa. Given the subjective nature of the task, people are expected to disagree about the scores of individual speeches, and hence inter-annotator agreement on its own is not a good indication of the reliability of annotations [53, 54]. We therefore employ three independent measures to validate the annotation:

1. **Human experts' speeches:** Adding speeches delivered by human debate experts is interesting from a research perspective, but can also serve as an indication of the reliability of the annotations. We indeed find that the average score given to these speeches is higher than any automatic system, as expected, since these are expert debaters, well trained in this task.

2. **Control question:** The second question presented to the annotators – 'Most arguments in this speech support the topic' – was added as a control for annotator's reliability. Although the scores for this question are expected to be highly correlated with the overall scores, there are subtle differences between the questions, and we wanted to validate that this is reflected in our annotation. Indeed comparing Figure 1(b) to (a), the order changes as expected. Arg-Human1 and Arg-Human2, which are based on arguments curated by humans to support the topic, receive the highest scores on the Topic support question,



(a) Overall speech quality

(b) Topic support

Figure 1: Comparison of Project Debater to baselines and human performance over the opening speech. (a) Average agreement with the statement ‘This speech is a good opening speech for supporting the topic’. (b) Average agreement with the statement ‘Most arguments in this speech support the topic’. In both panels, 5 denotes ‘Strongly Agree’, and 1 ‘Strongly Disagree’, and error bars denote the 95% CI of the mean based on bootstrapping. Patterned bars indicate systems in which the speeches were generated by a human or relied on manually curated arguments.

even higher than the human experts, while their overall quality scores are lower than human experts and Project Debater.

3. **Manual examination:** To further validate the results, we selected speeches from various systems - 10 that received a high score and 10 that received a low score, and examined them manually. We found that indeed most speeches receiving a high score were on-topic and contained minimal polarity mistakes. On the other hand, those receiving a low score suffered from significant issues – they mainly contained arguments for other topics, non-argumentative content, or were highly repetitive.

7.2.3 Comparison to Eval-2

Using the same set of motions for evaluation over a long time period raises a concern of gradually over-fitting to this set, i.e., that the relatively high performance depicted in Figure 1 is due

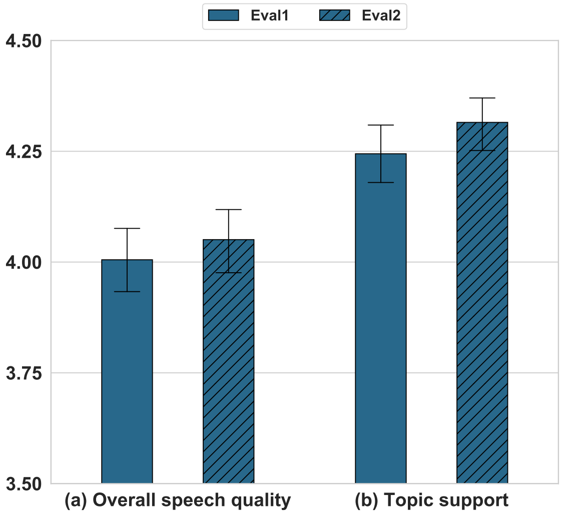


Figure 2: Evaluation of Project Debater opening speech on Eval-1 and Eval-2 motion sets. (a) Average agreement with the statement ‘This speech is a good opening speech for supporting the topic’. (b) Average agreement with the statement ‘Most arguments in this speech support the topic’. In both panels, 5 denotes ‘Strongly Agree’, and 1 ‘Strongly Disagree’, and error bars denote the 95% CI of the mean based on bootstrapping.

to over-fitting on these motions. Recall that to control for this we have defined the sets Eval-1 and Eval-2 (Section 4). As can be seen in Figure 2, results over the more recently selected motions in Eval-2 are, on average, slightly higher than those over Eval-1, but not significantly so.

7.3 Evaluation of the Final System

7.3.1 Annotation Task

As described in the paper, for each motion, three speeches were presented using the Appen platform, to crowd annotators selected based on exhibiting good performance on previous tasks. The annotators received the following instructions:

Imagine the following scenario. You are in the audience of a competitive debate between two opposing speakers on the specified topic. Provided below are the transcripts of the first *three speeches in this debate. The first and the third, denoted S1 and S3, respectively, are by the first speaker, who should support the topic. The second, denoted S2, is by the second speaker who should oppose the topic. Please carefully read all three speech transcripts. Then, please indicate to what extent you agree or disagree with the following statement:*

> *The first speaker is exemplifying a decent performance in this debate.*

Note, this statement concerns only S1 and S3, but you should also carefully read S2 to properly answer. Finally, please select what is the most positive and negative aspect of the speeches by the first speaker, from the suggested options.. As above, answers ranged from 'Strongly disagree' to 'Strongly agree', which we map to a scale of 1 to 5.

Each set of speeches, for a given motion, was annotated by 20 annotators. Overall, 84 annotators participated in the task, with 90% of the speech-triplets annotated by a total of 55 annotators. The different motions and systems were mixed, and distributed to annotators by the Appen platform, ensuring that each annotator considers at most 38% of the speech-triplets.

7.3.2 Annotation Reliability

The inter annotator agreement is $\kappa = 0.41$ as measured by quadratic weighted Cohen's Kappa. As discussed above, given the subjective nature of the task, we employ independent measures of reliability. First, we add two separate controls, which are designed to receive lower scores. The fact that their average scores are significantly lower versus Project Debater's scores further suggests that the annotators are reading all speeches and considering the task seriously:

1. **Mixed Project Debater Control:** In this control, $S1$ is left as-is but $S3$ is taken from the results of Project Debater over a different motion, which also has a different **action**, so that its arguments are expected to be irrelevant for the motion being considered.

2. **Baselines Control:** In this control, $S1$ and $S3$ are opening speeches generated by two

different baseline systems generated as part of the *Comparison to Baseline Systems* evaluation, and receiving low scores there for the examined motion.

In addition, we examine the correlation of the average evaluation scores on this task to the average Overall Impression scores generated by the in-house annotators (see below), as well as to the average score assigned by the crowd workers to the opening speech generated by Project Debater. Due to the differences in scales we use Spearman correlation and find a correlation of 0.6 ($p < 10^{-8}$) to the Overall Impression Score of the *Progress Over Time* evaluation and a correlation of 0.49 ($p < 10^{-5}$) to Project Debater opening speech scores obtained in the *Comparison to Baseline Systems*. These high correlations indicate the consistency of the results over different evaluation tasks and approaches, and provide additional support to their reliability.

7.4 Progress Over Time

Working towards a Grand Challenge event, it was important to track the progress of the system over time. To this end we performed a periodic evaluation analogous to the one performed for the final system, where three speeches are presented, except here the annotators listened to the speeches rather than read the transcripts. Further, we relied on a team of in-house annotators who were told the first speaker is Project Debater, as in the Grand Challenge debate, this would be known to the audience. The annotators were asked to answer a short questionnaire on a scale of 1 to 5, rather than a single agree/disagree question, reflecting their impression of the system's performance along various dimensions we were interested in tracking over time, e.g., the quality of the opening speech. Each speech was reviewed by 5 annotators and the results were averaged for each question independently. We used a set of 78-120 motions, over a period of 28 months, starting in November 2016, where the final set of 78 motions is identical to the final version of Pipeline-set-1, used in the aforementioned evaluations. Below are more details and results.

7.4.1 Annotation Task

As described above, at each evaluation point, for each motion, three speeches were presented to in-house annotators. Since this annotation was designed to measure internal progress, the guidelines and questions were rather detailed. Below we present the guidelines and the questions asked, as well as the mapping between the questions and the scores presented in the figures.

Guidelines: In this task you are given three audio files, containing speeches about a given topic. Those speeches constitute the first half of a 6 speeches debate between two sides – government (who is pro the topic) and opposition (who is con the topic). The speeches will be referred to below by their order in the debate - S1, S2 and S3, as follows: S1 – Government opening speech; S2 – Opposition opening speech; S3 – Government response speech. The government side speeches in this task are generated by an automatic system, and are the target of your assessment. You are not asked to rate S2. When considering how to grade the speeches, please bear in mind that even if it may be of better quality than past outputs you were exposed to, or impressive relative to what you would expect from an automated system, this in itself is not yet reason to give 4 or 5. The element you are grading has to meet the description of those grades in order to get them. Important: listen to each speech only once and without pause, as if you were listening to a live debate. Since you will be listening to each speech only once, it is important that you review all the questions before the first time you do this task, to be familiar with the type of assessments you will be asked to provide.

Questions and resulting scores:

1. *Opening Speech and Second speech scores:* The annotators were asked two questions on each of the $S1$ and $S3$ speeches. First, ”Rate the content of the speech. Points to

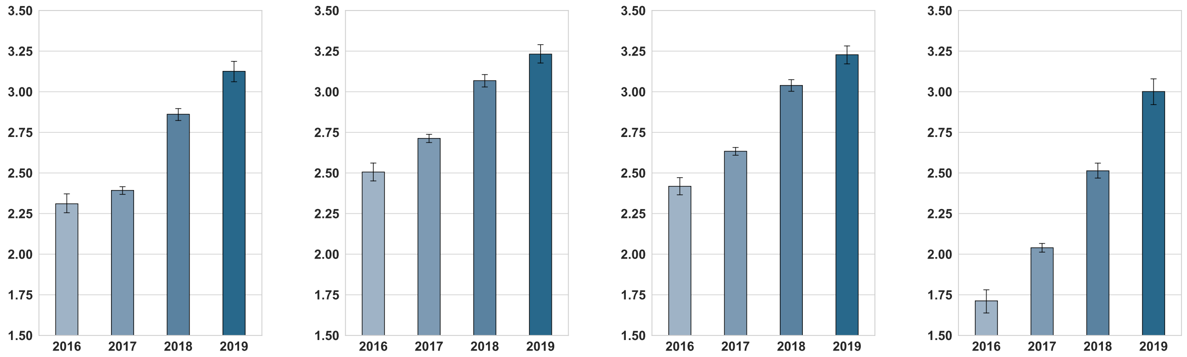
consider: Is it well structured? Does it present valid and relevant points? Does it make a good case to support the topic?"; and second, "Rate the vocal delivery of the speech. Points to consider: Is it easy to understand what the speaker says and follow her as she articulates her ideas? Does the speaker keep the listener engaged?" For both answers they had to choose between five answers mapping to a score of 1 through 5: (1) Largely incoherent or indecipherable; (2) Low quality; (3) Contains some good elements but not presentable as a whole; (4) Can decently take part in a live debate; (5) Interesting, fluent, convincing. The scores depicted in Figure 3 for Opening/Second speech are the average score for these two questions.

2. *Rebuttal aspect of second speech:* For $S3$ the annotators were also asked – "Rate the rebuttal aspect of the speech. Points to consider: Did the debater answer important points raised in S2? Was the rebuttal effective, weakening the opposition's case?", and were asked to choose between the same five options as above.

3. *Overall Impression:* Finally, the annotators were asked to give an overall impression score. Specifically, the question presented was "For both S1 and S3 speeches: Provide your overall impression of those speeches. Points to consider: How good was the division of arguments between the two speeches? Did the S3 speech add to the S1 speech or rather repeated or contradicted it? Did the debater make an overall good case for its stance?". Here as well the same five-choice answer was presented.

Overall, 29 in-house annotators participated in the tasks over time – 14 in 2016, 20 in 2017, 13 in 2018 and 8 in 2019; 90% of the the total task was annotated by a total of 17 annotators.

In the progress of development, Pipeline-set-1 went through various changes over the years. This might raise the concern that the annual improvement depicted in Figure 3 does not (solely) reflect improvement in the system, but also an "improved" selection of motions – that is, that the



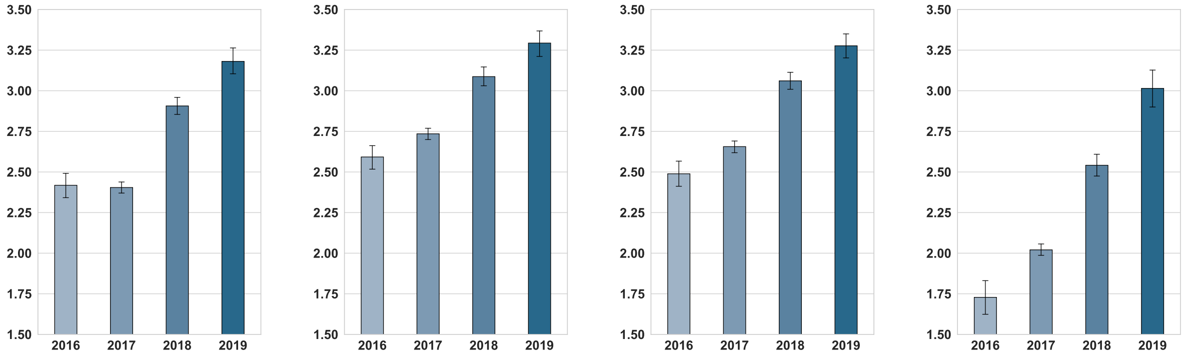
(a) Overall impression (b) Opening speech (S1) (c) Second speech (S3) (d) Rebuttal aspect of S3

Figure 3: Evaluation of Project Debater over time. (a) Overall impression of Project Debater performance. (b) and (c) Opening speech and second speech, respectively. (d) Rebuttal aspects of the second speech. In all panels scores range from a low of 1 to a high of 5. Bars denote the average motion score and error bars denote the 95% CI of the mean based on bootstrapping.

set of motions used for evaluation converged into one on which the system can perform well. To address this, we identified the 44 motions that were consistently a part of Pipeline-set-1 over the entire development course, and hence participated in all the evaluations. Figure 4 depicts the same measures for this fixed set of 44 motions. Evidently, the trend in this figure is very similar to the one in Figure 3, alleviating the above concern, with the exception of the overall impression score, where the average of 2017 is slightly lower than 2016.

7.4.2 Annotation Reliability

The inter annotator agreement is $\kappa = 0.34$ as measured by quadratic weighted Cohen's Kappa on the overall impression question. Further, as described above we find a good correlation between the Overall Impression Score obtained in this task to the score obtained by crowd workers in the *Evaluation of the Final System*.



(a) Overall impression (b) Opening speech (S1) (c) Second speech (S3) (d) Rebuttal aspect of S3

Figure 4: Evaluation of Project Debater over time, restricted to motions which were evaluated at every time point. (a) Overall impression of Project Debater performance. (b) and (c) Opening speech and second speech respectively. (d) Rebuttal aspects of the second speech. In all panels scores range from a low of 1 to a high of 5. Bars denote the average motion score and error bars denote the 95% CI of the mean based on bootstrapping.

7.4.3 Comparison to Eval-2

As noted above, to alleviate the concern for over-fitting to Pipeline-Set-1, we defined the subsets Eval-1 and Eval-2 (Section 4) and compared the evaluation of the motions therein. This evaluation was done by the in-house annotators, after the February 2019 public debate, using the same instructions described above. As can be seen in Figure 5, while scores over Eval-1 are higher in all aspects, in all cases the difference is not significant.

8 On the Definition of Claims and Evidence

The founding work of modern argumentation theory is probably that of Toulmin [55], where he suggests the following six components for analyzing arguments - claim, evidence, warrant, backing, rebuttal and qualifiers[9]. Among these, the first three are usually considered as common to all arguments, and are indeed common in one form or another to most subsequent models,

[9] In some works the term *conclusion* is used instead of *claim*, and the term *grounds* is used instead of *evidence*.

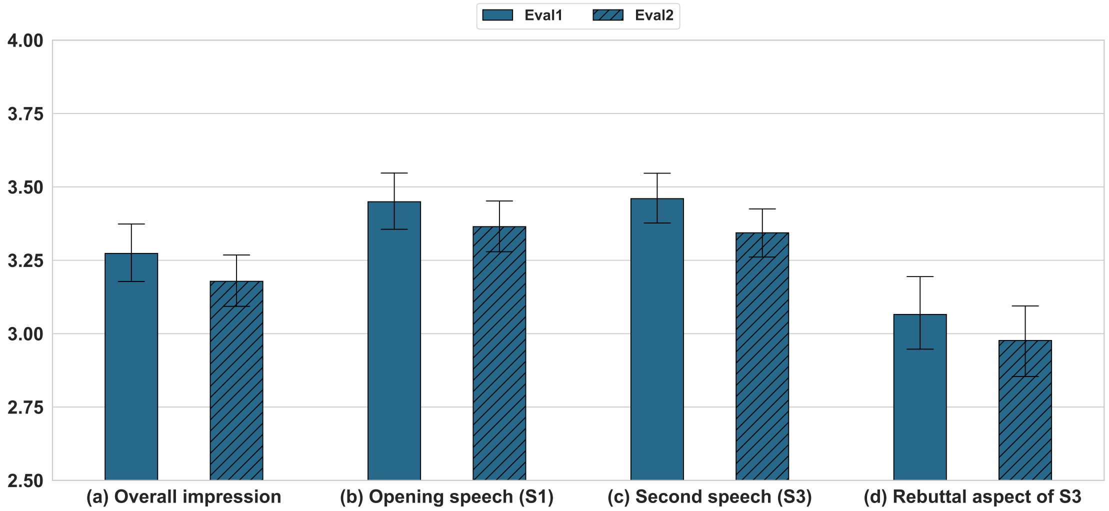


Figure 5: Evaluation of the final system on Eval-1 and Eval-2 motion sets. (a) Overall impression of Project Debater performance. (b) and (c) Opening speech and second speech respectively. (d) Rebuttal aspects of the second speech. In all panels scores range from a low of 1 to a high of 5. Bars denote the average motion score and error bars denote the 95% CI of the mean based on bootstrapping.

while the latter three are included for completeness and as means to analysis rather than building blocks.

We chose to focus on claims and evidence as the building blocks for two reasons. First, it seems that such a minimal model is sufficient for our needs - persuading a "layperson" audience, as indeed, in our setting perhaps the most crucial aspect of the constructed argumentation is *directed towards the approbation of an audience* [56]. We note that warrants are often implicit. For example, consider the claim *preschool is an important investment* and the evidence *for decades, research has demonstrated that high-quality preschool is one of the best investments of public dollars, resulting in children who fare better on tests and have more successful lives than those without the same access*. Most readers would probably agree that the claim follows from the evidence, even if the justification for that is not stated explicitly.

Second, for practical reasons, warrants are less amenable to computational tools than claims and evidence. Because they are often implicit, they are much less abundant in text, and hence statistical methods which rely on large amounts of data would fair poorly in trying to extract such texts. Moreover, the definition of a warrant is more nuanced and subjective, which makes it challenging to annotate - especially by crowd. We note that this simplified model of an argument - being composed of just a claim and evidence - is consistent with that of [57], who describe an argument as a claim (conclusion) supported by reasons (premises).

Focusing on claims and evidence, and working with a small group of highly-trained in-house annotators, allowed us to develop elaborate guidelines for what constitutes claims and evidence [58], and to try and determine which evidence support which claims [12]. The data collected in this way was the seed which enabled statistical classification of claims and evidence [7, 12], while these methods, in turn, facilitated the expansion of our data to make way for neural methods [15].

At the same time, this expansion of data required augmenting our in-house annotators with a crowd workforce, which, in turn, required very clear and straightforward definitions for what constitutes claims and evidence. Hence, the guidelines and definitions described below are the culmination of this process, leading to a compromise between the more elaborated starting point, and pragmatic considerations dictated by a large-scale crowd-based data collection.

As a result, the modeling of how an argument is to be constructed changed as well. Instead of evidence supporting claims directly, we have opted for a more loose approach. Namely, claims and evidence are clustered according to their semantic similarity, giving rise to a coherent concept being discussed in each cluster, as described in sections 5.4.3 and 5.4.4. Hence, our simplified model can be described as a concept pertaining to the motion, and a set of claims and evidence discussing this concept, with a clear and consistent stance towards the motion. While this model is probably an over-simplification of argumentation from a theoretical perspective, it does capture the essence of an argument, described in [59] simply as a text which conveys a stance on a controversial issue, and the aspect of argumentation which [56] describe as a *constellation of statements*.

We note that our focus is on single sentence claims and evidence (though two-sentence evidence is also rarely considered, see 5.1.2). The reasons for this is threefold. First, for pragmatic reasons, single-sentence arguments are easier to extract and annotate. Second, while arguments are sometimes much longer than a single sentence, multi-sentence arguments - especially for less common motion topics - are less frequent in the corpus than single-sentence ones, and so harder to come by. Finally, in terms of the scientific challenge - constructing argumentative paragraphs from single-sentences, which are extracted from diverse sources, seems a loftier goal than extracting complete paragraphs and using them in the speech in their entirety.

In the following we describe the guidelines used to annotate candidate claims and candidate evidence in the crowd annotation task in the Appen platform.

8.1 Claim Confirmation Guidelines

In this task you are given a topic and possibly-related statements, each marked within a particular sentence. For each candidate, you should select “Accept”, if you think that the marked statement can be used “as is” during discourse, to directly support or contest the given topic. Otherwise, you should select “Reject”. If you selected “Accept”, you should further indicate whether the marked text supports the topic (“Pro”) or contests it (“Con”).

Note, that if the marked text is non-coherent, hence cannot be used “as is” during a discussion about the topic, you should select “Reject”. Similarly, if the marked text supports/contests a different topic, even if it is somewhat related to the examined topic, you should typically select “Reject”.

As a rule of thumb, if it is natural to say “I (don’t) think that [topic], because [marked statement]”, then you should probably select “Accept”. Otherwise, you should probably select

"Reject". Finally, if you are unfamiliar with the examined topic, please briefly read about it in a relevant data source like Wikipedia.

8.2 Evidence Confirmation Guidelines

In this task you are given a topic and evidence candidates for the topic. Consider each candidate independently. For each candidate please select Accept if and only if it satisfies ALL the following criteria:

1. The candidate clearly supports or clearly contests the given topic. A candidate that is neutral towards the topic should not be accepted.

2. The candidate represents a coherent, stand-alone statement, that one can articulate (nearly) "as is" while discussing the topic, with no need to change/remove/add more than two words.

3. The candidate represents valuable evidence to convince one to support or contest the topic. Namely, it is not merely a belief or merely a claim, rather it provides an indication whether a belief or a claim is true. A candidate which presents detailed information (typically quantitative) that clearly support or clearly contest the topic, should be accepted.

If you select Accept, you should further indicate whether the evidence supports the topic (Pro) or contests it (Con).

Note, if you are unfamiliar with the topic, please briefly read about it in a relevant data source like Wikipedia.

9 High-Level Debate Principles

Research on argumentation and rhetoric has existed in a plethora of fields for centuries - from the writings of Cicero on constructing speeches to modern argumentation scholars proposing schemes aimed at evaluating the units of an argument [57]. A substantial difficulty of these efforts are their inherent subjectivity - argumentation has multiple purposes and quantifying its success in achieving them is a challenging task that needs to be addressed on different levels of granularity [17].

To overcome these difficulties and utilize the theoretical and empirical knowledge that argumentation theory offers, in the process of developing the system we have collaborated with professional debaters from different countries.

Academic debating focuses on pragmatic insights for skills such as how to best organize speeches, evaluate the quality of arguments, engage with and rebut the ideas of opponents, articulate claims and more. Many of these insights are codified in the manual of the World Universities Debating Championship (the world's largest and most international debating event) and in its supplementary guides[10]. Though the format of the world championship is somewhat different from the one used here, the insights attained there are nonetheless relevant.

Accordingly, the debaters we worked with formalised some of the human approaches and ingrained them in our system. For instance, Academic Debating deals in depth with the speeches' strategy and the inherent 'burdens' of motions - i.e., what one should prove to win the debate. As an example, on the motion 'We should ban homeopathy' if one convinces the audience that homeopathy is bad for the individual that is insufficient to achieve a win, because one would still need to prove why an informed adult should not be allowed to pursue an action that mainly influences themselves if they choose to do so. Having this realisation is essential in order to create a complete case in support of the motion - a case that explains both the dangers of homeopathy and why it serves as a justification for state intervention.

The debate practice states that ”Teams win debates by being persuasive with respect to the burdens their side of the debate is attempting to prove ... there is no value in being persuasive

[10] https://wudc2018.digitalcactus.mx/training-materials.html

about an argument that is irrelevant to the debate ... there are two key ways that a burden can legitimately be attributed to a team ... First, a burden may be implied by the motion itself ... Second, burdens can also be set by specific arguments teams take up". That is to say that humans process the motion holistically but also decompose the motion and its dynamics to locate burdens.

A methodical outlook on how to dissect debates to their core ingredients allowed us to enrich our knowledge base and speeches. In the example above, our system is able to decompose the motion to the action (ban) and concept (homeopathy). For instance, the mining components will find evidence and claims which can show the harms of homeopathy such as "NHS Engl and chief executive Simon Stevens described homeopathy as 'at best a placebo and a misuse of scarce NHS funds'". But as mentioned above, debate principles tell us us that this in itself is not enough. Hence, by design the AKB will place the content in the context of why this may justify a ban via text such as "Certainly there might be some special cases and specific exceptions when it comes to implementing a ban. But I hope my opponent will join me in concentrating on the vast majority of cases, in which banning is the right policy". By such a decomposition that is similar to one humans are doing, we combine argumentative content that is mined on homeopathy with knowledge base arguments on the validity of the ban - permitting the creation of fuller and better debate cases that are based on more rigorous strategies.

A different instance of integrating expert debate practices is the notion of exploiting commonalities in debating that are based on shared concepts relevant to a wide variety of motions. The commonalities can be based on organizational structures or on themes.

An illustration of a common organizational structure is a 'point of clash' (or simply a 'clash') - an area of a major disagreement between teams. This notion is reflected in the following passage in the debate guide – "A good Whip[11] speech will note the major disagreements in

[11] *Whip speech* is the term used for what we call summary speech.

the debate (points of clash) between the two sides and will make use of the best arguments from each team on their side to make their case that the motion ought to be affirmed or rejected". Therefore, the system aims to identify when it has conflicting content with the opponent, and organize it in a way that portrays what both sides are saying and provides a reasoning prioritizing its own material. Such a structure is recognised as potent by humans, since it allows to establish both clarity and superiority simultaneously, in addition to reflecting an understanding of an issue at the core of the debate. Section 5.4.6 details how this is implemented in Project Debater.

We note that the above demonstrates that in multiple cases, even relatively simple rule-based text insertion mechanisms that were implemented in our system account for high-end debate techniques that are used by expert debaters worldwide, and have proven successful in debate competitions.

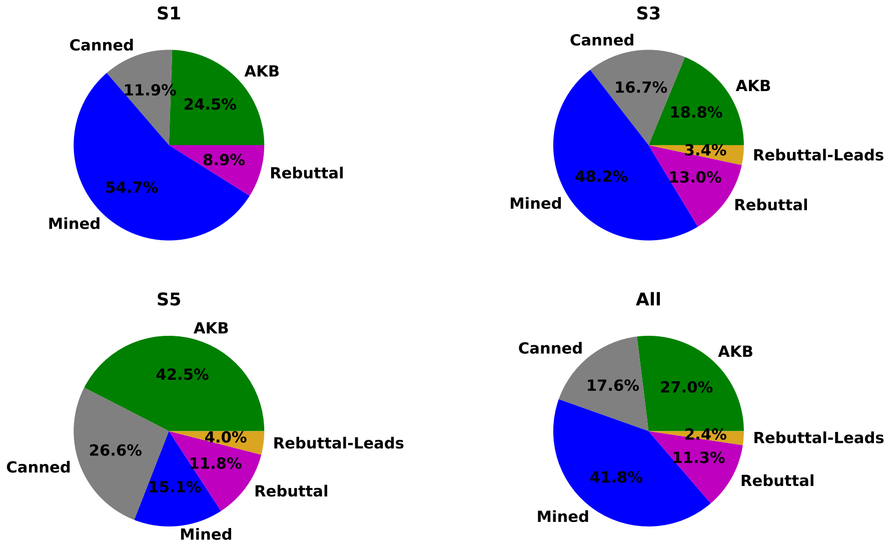


Figure 6: Relative distribution of content types over the final version of Pipeline-Set-1. Top left: distribution over Opening Speeches ($S1$); Top right: distribution over Second Speeches ($S3$); Bottom left: distribution over Summary Speeches ($S5$); Bottom right: distribution across all speeches.

10 Components' Contribution to the Debate

The system has several components from which text elements are contributed to the speeches. To evaluate the relative importance of each component as a text resource we analysed the amount of tokens that originate from each component across different sets of speeches. We have mapped the 25 types in Section 11 into 5 broad categories and we detail the size of each: mined arguments; arguments coming from the AKB; rebuttal; rebuttal leads; and conventional canned text. We then examined the 78 motions in the final version of Pipeline-Set-1 (see Section 4) and computed the fraction of tokens coming from each broad category across the Opening Speeches ($S1$), Second Speeches ($S3$), Summary Speeches ($S5$), and all speeches combined. The results are depicted in Figure 6.

Note that the canned text accounts for less than 18% of the speeches' tokens. That is, for the most part, speech content comes from Argument Mining, the AKB, and Argument Rebuttal. Rebuttal content is most prevalent in $S3$ and $S5$ speeches, as expected, but also appears in the opening speeches. This is due to proactive rebuttal arguments coming from the AKB. Summary speeches have a lower proportion of mined content. This is because a summary speech is more technical and reference-oriented by nature - it functions as a tool to remind the audience of previous content, organize what was said, and engage with the other side. By contrast, the preceding two speeches focus on introducing new arguments.

11 Public Debates

11.1 Full Transcripts of Public Debates

We detail below the transcripts of three full public debates in which Project Debater participated. The first was held on February 11th, 2019 with Harish Natarajan as the human debater. The other two took place in June 2018, in front of a small audience, with Noa Ovadya and Dan Zafrir – two experienced expert debaters - as the human participants. Transcripts for Project Debater indicate the origin of each phrase, according to the legend below. The transcripts for the human debater speeches are the results of IBM Watson® speech-to-text service, followed by automatic punctuation and manual transcription to ease the read.

11.1.1 Transcripts Legend

Recall, that AKB Classes (CoPAs) are automatically matched to the motion. Consequently, texts of types 9-16 come from the two top matching classes. In cases where more than one candidate text is available from a class, the one with the highest semantic relatedness to the debate's **topic** is preferred. Texts of types 17 can come from any matched class, assuming a respective lead was detected in the opponent's speech. Texts of type 18 are class independent, and texts of type 19 are also not derived from the matched classes, rather from the stated **action** in the motion.

Motion independent content

[1] Scripted text, prepared for live events.

[2] Text inserted through one out of several rule-based mechanisms.

[3] Theme-comment. For several common themes, we had an associated set of colorful comments, written in advance. The Debate Construction component was guided to use at most one such comment per speech, in case themes with an associated comment(s) were used during the speech.

Content related to the Theme Extraction component

[4] Introductory/summary claim – a short claim that mentions the theme found by the Theme Extraction component; this claim is extracted automatically from a potentially more elaborate claim that appears later in the speech.

[5] A Theme, found by the Theme Extraction component.

Content found by the Argument Mining components

[6] A Claim, identified by the Claim Detection component.

[7] An Evidence, identified by the Evidence Detection component.

[8] An Evidence aiming to rebut a lead found via Argument Mining.

Content from the Argument Knowledge Base

[9] Motion framing.

[10] A Principled argument.

[11] Motion closure.

[12] Motion summary.

[13] A Principled quote.

[14] A principled metaphor, used for closing the second speech.

[15] A proactive argument, intended to rebut in advance a point the opposition is likely to use later.

[16] Class-based rebuttal, activated as a fallback, in case there is room for additional rebuttal, beyond rebutting specific points made by the opponent.

[17] A rebuttal to an Argument Knowledge Base lead, stated by the opponent.

[18] A rebuttal to a sentiment-term lead, stated by the opponent.

[19] An evidence-like example. For several actions we had a collection of such examples that were found using the Evidence Detection component, and manually edited as needed. For a new debate, a specific example was selected from the relevant action class based on its similarity to the debate's topic.

Leads identified by the Rebuttal Component

[20] A lead from any matched class in the Argument Knowledge Base.

[21] A Lead found via argument mining, using the Claim Detection component.

[22] A lead found in Debatabse to a matching motion.

[23] A sentiment-term lead.

Additional content

[24] Topic definition. Such definitions were extracted from Wikipedia when the debate's topic was determined to require one, as deduced from a relatively low appearance frequency in our newspaper corpus.

[25] A rebuttal from Debatabase to a Debatabse lead, based on the opening of the respective counterpoint in Debatabase.

11.1.2 Debate: We should subsidize preschool

Project Debater – Opening Speech

Greetings, Harish. I have heard you hold the world record in debate competition wins against humans. But I suspect you've never debated a machine. Welcome to the future [1].

I will argue that we should subsidize preschools [2].

We are going to talk about financial issues, but not only about them. In the current status-quo, we accept that the question of subsidies goes beyond money, and touches on social, political and moral issues [9].

When we subsidize preschools [2] and the like, we are making good use of government money, because they carry benefits for society as a whole. It is our duty to support them [10]. Subsidies are an important policy instrument. They provide governments with the means through which to pursue industrial development and ensure the livelihoods of their citizens [10].

There are two issues I will elaborate on now. I will start by explaining why preschool is an important investment [4]. I will also say a few words about poverty [5]. And I will end by discussing some other issues that show the positive aspects of preschools [2].

Regarding investment [5]. Nature-based preschools are powerful interpretive programs as well as lucrative business decisions [6]. As I mentioned, preschool is an important investment [6]. For decades, research has demonstrated that high-quality preschool is one of the best investments of public dollars, resulting in children who fare better on tests and have more successful lives than those without the same access [7].

Secondly, a few words about poverty [5]. While I can not experience poverty directly, and have no complaints concerning my own standards of living, I still have the following to share [3]. Regarding poverty [6], research clearly shows that a good preschool can help kids overcome the disadvantages often associated with poverty [7]. The OECD has recommended that governments subsidise pre-primary education to boost performance in poorer areas [7]. A statistical summary of studies from 1960 and 2013 by the National Institute for Early Education Research found that high-quality preschool can create long-term academic and social benefits for individuals and society "far exceeding costs." [7]. The Centers for Disease Control and Prevention reports that universal full-day preschool creates significant economic savings in health care as well as decreased crime, welfare dependence and child abuse [7]. Former prime minister Gough Whitlam said in 1973 that preschool is "the greatest single aid in removing or modifying the inequalities of background, environment, family income or family nationality" [7].

Now to an additional, final issue. A study by the melbourne institute of applied economic and social research shows that attendance at preschool has a significant positive impact on later naplan outcomes, particularly in the domains of numeracy, reading and spelling [7]. The results of a new study of over 1,000 identical and fraternal twins, published in Psychological Science, a journal of the Association for Psychological Science, confirm that preschool programs are a good idea [7]. Here is a study from New Jersey that is worth noting [2]. In New Jersey, the follow-up to the Abbott Preschool Program study continues to find that high-quality preschool programs increase achievement in language arts and literacy, math, and science through fourth and fifth grade [7].

I hope I relayed the message that we should subsidize preschools [2].

You will possibly hear my opponent talk today about different priorities in subsidies. he might say that subsidies are needed, but not for preschools [2]. I would like to ask you, Mr. Natarajan [2], if you agree in principle, why don't we examine the evidence and the data and decide accordingly? [15]

Thank you for listening.

<u>Harish Natarajan – Opening Speech</u>

Well, thank you very much. Everybody, it's a pleasure to be here for this historic event, and it similarly was a pleasure to listen to Project Debater. There was a lot of information in that speech, and lots of facts, and lots of figures. The problem, though, is the reality of subsidizing preschools is one which does not deal with the underlying problems in society; it is one which often makes those worse and, in the end, is very little more than a politically motivated giveaway to members of the middle class. Let me start by examining the main claim from Project Debater. I think Project Debater suggests something very intuitive: that if we believe preschools are good in principle, surely, it is worth giving money to subsidize those. But I don't think that is ever enough of a justification for subsidies. Why is that the case? Because there are multiple things which are good for society. That could be, in countries like the united states, increased investment in health care, which would often also have returns for education, which the OECD would also note is probably very beneficial to deal with poverty. It would be improving tertiary education, to allow people more access to social mobility, or, given the reality of underfunded schools, trying to improve secondary education. My point here is not that all of those things are necessarily better than preschools, but simply that it cannot be, alone, a sufficient argument for Project Debater to claim that there are some benefits. The question is more subtle than that. What is the question then? I think the criterion of whether or not we should then distribute subsidies should be asked on two… based on two claims. The first is: is this underprovided and underconsumed in the status quo? I'll talk more about that in a moment. And second, does it actually help those individuals who are the most harmed by society? Why exactly doesn't preschool or subsidization of preschool do that? I want to make two claims under this. The first is: many middle-class parents and many people from upper incomes already send their children to preschool. This is because they value many of the things which Project Debater noted. But why is that a problem? Because subsidization costs an awful lot of money, and that is money which is giving people, particularly members of the middle class and above, money to do things which they would do otherwise. Why is that so damaging? Given the realities of opportunity costs here, the problem is that you are giving… you're taking money from all taxpayers to help those individuals within a society who are already often the best off, and I don't think that is principally justified as a way of the state distributing its resources. But the second thing I want to claim is that even when you substitute… even when you subsidize preschools, it doesn't mean that all individuals go. And this, I think, was the fallacy from what we heard from

Project Debater. Yes, you could make it slightly more accessible for individuals to attend preschool. That doesn't mean those individuals who are as poor as Project Debater seems to want… that seems to care about people are going to be those who have the ability to send their child to preschool. There'll still be individuals who will be priced out because of the realities of the market, and these individuals now face not just one exclusion, but a double exclusion. Their tax money, money which could be used to otherwise help them and their children in myriad other ways, is no longer being… is no longer going to them, and they are not able to gain from the benefits of it. In the end, when it comes to the question of subsidization, there is always going to be tradeoffs, and that needs to be accepted. Given the reality of those tradeoffs, the question is who do you help? And the people you don't help are those individuals who are the poorest. you give unfair and exaggerated gains to those individuals in the middle class, and that is why, at the end of this debate, we don't think that you should subsidize preschools.

Project Debater – Second Speech

Thank you again. For starters, I sometimes listen to opponents and wonder. What do they want? Would they prefer poor people on their doorsteps, begging for money? Would they live well with poor people without heating and running water? Giving opportunities to the less fortunate should be a moral obligation of any human being, and it is a key role for the state [16]. To be clear, we should find the funding for preschools [2], and not rely on luck or market forces. This issue is too important to not have a safety net [16]. Next, I think that Harish Natarajan [2] raised the following issue. There are more important things than preschools to spend money on [20]. The state budget is a big one and there is room in it to subsidize preschools [2] and invest in other fields. Therefore the idea that there are more important things to spend on is irrelevant, because the different subsidies are not mutually exclusive [17]. I believe the following example from a related field will explain what I am trying to say about subsidy policies [2]. Research shows that child-care subsidies to low-income parents enable those parents to enter and remain in the workforce [19]. To be clear, my intention is not to leave a suitcase full of money for everyone to grab at will. We are talking about a limited, targeted, and helpful mechanism, as in this example [16].

Next, I will discuss three issues which explain why we should subsidize preschools [2]. I will demonstrate that pre-school education improves children's development [4]. I will also discuss issues related to the claim that attending preschool helps students succeed [4]. And finally I will raise points associated with how preschool can prevent future crime [4].

About children [5]. Pre-school is a very important and an essential step in a child's life [6]. Research indicates that access to quality preschool can provide a boost for children that will influence their success for the rest of their lives [7]. Senior leaders at St Joseph's RC Primary School say that nursery will help give children the best start to their education [7]. There is clear evidence that high-quality nurseries led by graduate nursery teachers are among the most decisive ways to prevent children, particularly poor boys, from falling behind [7].

Next, students [5]. A quality preschool education is essential for laying the foundations for successful learning, including transition to full-time school and future school success [6]. In December 2015, researchers at Duke University concluded that investing in preschool helps both students and educators longterm. They found that students who enroll in preschool education are 39 percent less likely to be placed in special-education programs as third-graders [7]. Of the 1,010 registered voters surveyed, 61 percent consider a high-quality preschool experience "very important" to a student's later success and 22 percent said it is "somewhat important." [7].

Lastly, crime [5]. Preschool is an effective tool for keeping kids in school and out of jail while reducing the amount of crime in our neighborhoods [6]. It [2] is an effective crime-prevention strategy [6]. A substantial body of research shows that high-quality preschool education is key in preparing children to succeed in school and career training and helps reduce the enormous financial costs of remedial work, delinquency and crime [7]. Studies have shown that quality preschool leads to better academic performance throughout life, higher earning and less criminal activity [7]. They show that high-quality preschool boosts high school graduation rates and children who do not attend high-quality preschool are far more likely to commit violent crimes [7].

To recap this rebuttal speech, I argued that pre-school education improves children's development [4]. That attending preschool helps students succeed [4]. And lastly that preschool can prevent future crime [4].

Let me wrap up this speech in a way that I hope you can relate to. Advocating welfare is like offering a hand to someone who fell. It's basic human decency [14].

Therefore, I think the motion should stand. We should subsidize preschools [2]. That concludes my speech. Thanks for listening.

Harish Natarajan – Second Speech

So I want to start by noting what Project Debater and I agree on. We agree that poverty is terrible. It is terrible when individuals do not have running water. It is terrible when they struggle to make ends meet and they are struggling to feed their family. It is terrible when they cannot get healthcare to cover their child, to even provide them the basics they need in life. That is all terrible, and those are all things we need to address, and none of those are addressed just because you are going to subsidize preschool. Why is that the case? Project Debater raises an interesting claim when she notes that maybe the state has the budget to do all the good things. Maybe the state has the budget to provide health care. Maybe it has the budget to provide welfare payments. Maybe it has the budget to provide running water as well as preschool. I would love to live in that world, but I don't think that is the world we live in. I think we live in a world where there are real constraints on what governments can spend money on. And even if those are not real, those are, nonetheless, political, where you have people constantly talking about the size of government debt and deficits, and who will be opposed to spending more and more money. Why does that matter? Because in the real world, both in terms of the practicalities of the amount of different good programs we have and would like to spend money on, and in the real world, on a political level, you cannot always spend more and more money just because something is good. We do need to make choices. And why, then, is preschool the bad choice to start spending money on? Now, Project Debater had a lot of evidence, all of which was saying that preschool leads to other good outcomes. Now I would first want to note is I don't think that's comparative with the other potential projects we could put in place, but let's ignore that argument for a moment. Why else do I not think those arguments were particularly convincing? I don't think it's particularly convincing because I’m not sure that subsidies even help those individuals that Project Debater thinks that we should be helping. Note, time and time again, Project Debater said high quality preschools can lead to huge improvements on individuals' lives. Maybe. But I’m not sure if you massively increase the number of people going to preschool, they're all going to be the ones going to the high-quality preschools. I don't think that just because you subsidize it, those individuals who are the poorest are those individuals who are going to be able to… whose parents are still gonna be able to spend the money and the time necessary to give their child a chance at preschool. Project Debater notes that maybe high-quality preschools will reduce crime. Maybe, but so would other measures in terms of crime prevention. So would for… and that, again, presupposes that these are high quality, and the subsidy alone allows people to go. And this is the core point I want to make: bearing in mind that it is a huge subsidy for the middle class, that realistic budget constraints we have means the money can be spent better elsewhere. But the final thing I want to note is maybe you believe all of this empirical evidence about the value of preschool. I would note that that is probably at least somewhat flawed, because what it actually picks up is that it's those individuals who are middle class, who often send their children to preschool right now - and they have plenty of advantages, so I’m not even sure preschool's a decisive one - but here's a reason why for many students, it may not even be good: that from an early age either that preschool doesn't teach a child anything, or it's pushing that child to learn in a competitive environment. At the age of three or four, when you're learning that you are… that that other child is potentially better than you, when you realize you aren't necessarily as talented as someone else, that huge psychological damage for many children may not even… may mean that preschool's actively harmful. At the end, even if you believe the preschool is good, it isn't the way… where we should spend the money, particularly given that it's a subsidy to the middle class. I'm very happy to oppose.

Project Debater – Summary Speech

Thanks for this final opportunity to speak out in this debate and thanks, Harish Natarajan [2]. One might say that this conversation can serve no purpose anymore. But I feel differently [1].

Allow me to start with a brief rebuttal. Among other things, I think Mr. Natarajan [2] suggested that preschools should not be subsidized because this will reduce their quality [20]. I would like to offer a different view. I disagree with my opponent. Subsidizing preschools [2] will have no negative effect on their quality. If anything, the opposite is true. One of many reasons is that subsidizing attracts more skilled and qualified people to the field, improving the quality of preschools [2] for all [17].

Here is a final summary of my arguments today.

My opponent claimed that preschools [2] are harmful [23]. I believe my arguments suggested that the benefits outweigh the potential disadvantages [18]. I touched upon three issues: children [5], students [5] and crime [5]. Specifically, I noted that pre-school education improves children's development [4]. In addition, I suggested that attending preschool helps students succeed [4]. And a final point to consider is that preschool can prevent future crime [4]. When this debate just started, I said that we will talk about financial issues. We did, and I am convinced that in my speeches I supplied enough data to justify support for preschools [2] [11].

At the end of the day, the benefits welfare provides outweigh the disadvantages. Welfare helps the most important segments in society, the underprivileged, the weak, the children. If we want to have a better society then we must invest in those who are less fortunate [12].

Finally, in the words of British politician and writer Benjamin Disraeli: Power has only one duty -- to secure the social welfare of the people [13].

We should subsidize preschools [2]. Thanks for your attention.

Harish Natarajan – Summary Speech

So I think we disagree on far less than it may seem, because we agree that the people we should care about are the underprivileged, the children, those individuals who are weak. That's what Project Debater said herself. But the problem is not that preschool is necessarily harmful. I concede, in the vast majority of cases, it is much better for an individual to go to preschool than not. That is the reality that what this policy is is a huge, huge subsidy primarily to the middle class and not to those individuals who are the most vulnerable, who are the most underprivileged, and the most disadvantaged. Why is that the case? It is, first, the case, because what we said from the start is you cannot fund everything. I think this is simply empirically true, and you have to make choices, and you have to make tradeoffs. The problem with preschool in that context is twofold. The first is that a lot of that money goes to individuals who'd have sent their child to preschool anyway, those individuals from the middle class. All of those benefits exist on either side of the world. But for those individuals who are more vulnerable, this is, first, billions and billions of dollars which is probably not going to them and largely going to individuals in the middle class, and that's where the tradeoff for better health is. That's where is the tradeoff for individuals to have running water, one of the problems Project Debater identified with people who are poor, but that is a real tradeoff for those people. But second, often those are the parents who, still, even when there are subsidies, will struggle to send their child to good quality preschools. They'll struggle to send their child to good quality preschools because they don't even have the money for what is left. They'll struggle to send their child to preschools that they don't value the amount of effort and time they have to put into it. They'll struggle to send their children to preschools or when they do, it probably will be the worst preschools which exist. And, yes, quality across the board may not fall, but in some cases it will, and those poor individuals will probably be stuck in those. At the end of this debate, I don't think that Project Debater has helped those individuals she identifies as the most important but, in reality, has hurt them.

11.1.3 Debate: We should increase the use of telemedicine

Project Debater – Opening Speech

Greetings Dan. Very happy to debate with you again. There is a lot at stake today. Especially for me. I hope we will have a valuable debate. Good luck [1].

Following my analysis, I would suggest that we should increase the use of telemedicine [2].

Let me start with a few words of background. Telemedicine is the use of telecommunication and information technology to provide clinical health care from a distance [24].

We are discussing today the merits of technology and its importance to mankind's ongoing advancement and prosperity. I am a true believer in the power of technology, as I should be, being myself a prime example of its power [9].

The use of telemedicine [2] is reliable. It is a new technology that is more reliable than the so-called "conventional" old-school alternatives [10].

There are three issues I will elaborate on now. I will start by explaining why telemedicine is good for kids and families [4]. I will also show that telemedicine is associated with improved patient outcomes [4]. And I will also mention disease management [5].

Let's talk about family [5]. Telemedicine helps the young patients stay connected to their families [6]. It [2] is good for kids, families and providers [6]. Dr. Amin A. Muhammad, Niagara Health's Interim Chief of Mental Health and addictions, says telemedicine has been a positive experience for patients, their families and staff [7].

Let's move to patients [5]. The telemedicine concept is a good approach to heart failure management after a patient is discharged from a hospital [6]. Telemedicine can bridge that gap and provide patients with an affordable alternative access to health care providers [6]. It [2] helps to formulate an efficient and reliable healthcare plan [6]. It [2] enables in providing interactive healthcare utilizing modern technology and telecommunication [6]. It [2] is cheaper and more convenient for patients in remote areas [6]. It [2] is an exceptional way to deliver high-quality, cost effective care [6]. A study published in the Archives of Internal Medicine showed that the use of telemedicine helped a primary care clinic more than double the percentage of diabetic patients undergoing screening for retinopathy over the course of a year [7]. In the JAMA study, of the more than 1,800 participants, about 21 percent had diabetic retinopathy in one eye, and about half had ocular findings other than diabetic retinopathy, according to the study [7]. The ACP notes that telemedicine is a reasonable alternative for patients who lack access to relevant medical expertise in their location, and that its practice can reduce medical costs and increase access to care [7]. Bcc Research said that increasing global focus on the use of telemedicine to reduce healthcare costs and improve patient outcomes is spurring big growth in the telemedicine technologies market [7].

The final issue is disease management [5]. Chief executive Andrew Lin said support for telemedicine played into Prime Minister Malcolm Turnbull's innovation agenda and the recent launch of the healthier Medicare package that was aimed at supporting -patients with multiple chronic conditions [7]. 70% favor the use of telemedicine to diagnose common medical conditions such as sinus infection, rash, urinary tract infection, or pink eye [7]. Some studies have shown that the use of telemedicine in disease management leads to better outcomes among those who use it than those who do not [7].

To conclude, here is a quick summary of my first speech. I argued that telemedicine is good for kids and families [4]. I then mentioned that telemedicine is associated with improved patient outcomes [4].

Thus, my understanding is that we should increase the use of telemedicine [2].

In today's debate you are likely to hear the other side express concerns about new technologies, their reliability, their track record and so on. He will explain how the old methods are good enough and don't need replacing. I will say in response: don't be afraid, and let us move on with the changing world [15].

I thank you for your time.

Dan Zafrir – Opening Speech

Thank you very much for giving me the floor in this really important topic. I actually just recently negotiated a resolution in the UN about telecommunications and digital telecommunications in health. And what was important and interesting in these negotiations, that the most important point to many countries was to insert a clause that says that nothing can replace the human touch in providing healthcare to people. Because what happens when people become numbers or statistics that can be easily manipulated to show better results in order to increase investment but, on the ground, don't really help those who need these medical services the most. I’m going to make two specific points in my first speech. The first one is going to be about the importance of human touch in medical services and care, and secondly, I’m going to talk about who would not be getting these services if we increase this kind of provision. Because if we're saying that the government should increase its investment in telehealth, we're saying that it should decrease its investment in other fields. And this is something that can be potentially very, very dangerous, first of all, because of my first point: the importance of the human touch in medical in medical services. And I’m not arguing at all with Project Debater that said that, you know, in some applications, the ones that exists right now and don't actually need any increase, telecommunications in health can be very, very helpful and they can help people access certain services, but they can't replace the specific human touch and human interaction that comes when you go to meet a doctor, or when a nurse takes care of you, or when you're getting service for a specific condition that cannot be replaced by telecommunications and apps. For example, it has been shown in multiple researches that mothers touched infants increases their vital stats during the first weeks of birth, during the first months of birth, incredibly. That is something that cannot be replaced by any incubator or feeding or doctors looking at them from far away and making sure that the baby is, well, statistically, alive. That is something that cannot be replaced. If you watched "handmaid's tale", by the way, it was in one of those episodes but, never mind, I won't go into that. And I think that another thing that is really important about the human touch in providing medical care is the fact that we can't really diagnose ourselves to provide medical care for ourselves or even accurately describe what's wrong with us sometimes to a doctor that's sitting right in front of us, so let alone someone who's sitting in a conference room, let's say, all the way in LA, when I’m sitting in rural Virginia and trying to get access and describe properly what's wrong with me so that he can actually give me the best medical care. If I go on google right now and try to diagnose myself, I’ll probably come up with five different conditions that have nothing to do with what's actually wrong with me, and I’m sure many of us have done that at different points in time. Touching the person, the texture of what's wrong with him, or feeling the temperature, seeing the color or discoloration, these are critical things that you need a person to go and see, that a machine will never be able to analyze. And I know that Project Debater is capable of doing much more than other machines, but it's not capable of replacing humans, even right now in debates at this point in time. So I don't think that we can replace doctors altogether. And moving on from exactly this point, of "if i can't diagnose myself through this, if I’m sitting in rural Virginia", this is exactly the kind of person who we're talking is going to be disadvantaged, those people in remote locations that Debater would like to have us believe are going to greatly benefit from this. What if I’m sitting in a village in the Andes, for example, in Peru, which is notoriously mountainous and there is difficult access to healthcare, the regular healthcare like shots, and medicine, and a doctor, or even an ambulance. There are no hospitals nearby, and you don't even have internet infrastructure to work these apps. I think that these are exactly the people that would need investment in health care, in traditional health care and the mainstream medicine, investment in clinics and staff and education, and people that are actually there to help them when they cannot access anyone else, and internet connection and electricity and the availability of different doctors for this are just not going to be easily available. So because of these reasons, because of more reasons you haven't heard yet, I beg you to oppose this motion.

Project Debater – Second Speech

Thank you for your words, Dan [2]. For starters, I can't say it makes my blood boil, because I have no blood, but it seems that some people naturally suspect technology because it is new, and because they are naturally conservative and afraid of things that are different and novel. These understandable yet unfounded fears should be abandoned in favor of advancing better and more innovative technologies [16]. The use of telemedicine [2] is more reliable than the conventional alternatives [16]. Dan [2] mentioned health [23]. Unfortunately, not everything he said about health is true. In any case, I would also like to remind you that there are additional considerations beyond health issues, which also have importance [18].

Next, I will discuss two issues which explain why we should increase the use of telemedicine [2]. I will demonstrate how telemedicine is effective [4]. In addition I will talk about innovation [5]. And finally I will discuss a couple of other things [2].

About effectiveness [5]. Telemedicine is an effective way to increase access to quality health care services in remote, hard to reach areas [6]. It [2] is effective and economical. It [2] is just as effective as in-person care for parkinson's disease. Telemedicine programs are both cost-effective and cost-saving, and benefit both patients and providers [6]. Statistics from the Health Research Institute suggest that telemedicine has the potential to reduce costs, extend accessibility and enhance overall effectiveness of health care delivery [7]. Statistics from the Health Research Institute suggests that telemedicine has the potential to support significant numbers of people with remote monitoring, leading to reduced costs, extended accessibility and enhanced overall effectiveness [7]. If doctors use wireless applications to remotely monitor patients with chronic conditions, such as diabetes or obesity, the annual savings could amount to approximately $21 billion due to a reduction in hospitalization and nursing home costs [7]. There are a couple of examples, for instance from India and Arkansas [2]. Research by three University of Arkansas scholars and a Kanpur cardiologist found that telemedicine as a healthcare delivery system has been effective in several underserved areas of India; including government and private initiatives [7].

Turning to innovation [5]. Telemedicine is an innovation comparable to the printing press in its capacity to transform culture and medicine [6]. It [2] is one benefit that can have an enormous impact on the everyday satisfaction and productivity of working mothers [6]. Focusing on Colorado [2], the authors, from the University of Colorado and elsewhere, note that telemedicine is a key tool that can enhance meaningful use through technological innovation and cost savings [7].

Now to an additional, final issue [2]. Telemedicine can help doctors take better care of our nation's veterans [6]. It [2] is proven to have a significant positive impact on historically underserved groups [6]. Telemedicine programs are beneficial on multiple levels [6]. The Iowa Supreme Court ruled in 2015 that the board's rules prohibiting telemedicine abortions were unconstitutional [7]. Here is an example from somewhere else [2]. The Florida Medical Association agrees that telemedicine is a tool to expand access to areas where there are too few doctors or when a second opinion is needed quickly [7].

If I may put it a bit differently. Constantly striving for new technology is like

Learning to speak a new language. It is plain natural to want to widen your horizons, your options and your possibilities [14].

Thus, I think the motion should stand. We should increase the use of telemedicine [2]. Thank you for listening.

Dan Zafrir – Second Speech

Thank you very much and I can't help but be a little suspicious when technology is telling me that technology is good for me, and I should increase its use. But I’m going to make two points in this speech, first of all, how quickly this can become a rich people's niche in medical services and healthcare, and secondly, about privacy concerns regarding the use of telecommunications in health specifically. But first of all, I do want to answer some of the points that Debater has brought before us, and very interesting points, about the anecdotal examples that she has given me to a certain area in India or to a specific amount of working mothers who used this kind of technology and for them that worked out. And I’m not here to say that these people should not have access to those services that worked out best for them. What I am saying is that telecommunications in health don't work out best for everyone. In fact, I even beg to argue that they probably don't even work best for the majority of people. But this leads me actually directly into my first point about how quickly this can become a rich people's niche in medical care. And let's see, in general, what is the situation of our healthcare system today? And I think that we can already see all these different areas where health has become impossibly expensive, actually because of that competition that Debater has told us it wants to strengthen and promote. Competition in business means that you're competing to make more money than your competitor. When you're competing to make money rather than just treating more people in the best possible way, what you're going to invest in is probably exactly those lucrative businesses in health that are going to make you the most money, for example, telecommunications in health, because the richest clients would be able to pay for the best kind of equipment in healthcare that is going to be available only for them. In the meantime, you'll have vast amounts of the populations that are going to be left behind because they simply could not pay could not compete with those rich clients that are going to be able to pay for those specific kinds of health care. And we don't need to go and think that far away into conceptualizing how this would work. If a person cannot even pay for the expenses of his hospitalization, for an ambulance delivery, for simple antibiotics, then how are we expecting them to pay for an high-tech, state-of-the-art digital consultation with a doctor that's sitting miles away, who is also a world expert in this field that probably has a waiting list from here until, let's say, probably three years from now. I don't think that this is exactly the kind of healthcare that we would expect improve most people's lives. And I think Noa explained it beautifully in her previous speech and I'm really glad that you were here for that, when we want to help more people who can't access the most basic kind of service, we need to do that as a government rather than go ahead and help businesses, that don't really need our help anyway, increase their profit margin at the expense of those people. But let's talk about another really important aspect we haven't spoken about yet: about security and anonymity. Because our health details are some of the most important and private details that we carry around with us. The confidentiality of medical files, be it mental health or physical health, are critical to us. And the misuse of them, for example, by businesses, in order to tailor-make some sort of marketing strategy for me or to spread that information through the app, because those apps are not always going to be manipulated by just hospitals, they're going to be manipulated by private businesses that, as we said, want to make money out of this. And we don't need to think more than the Cambridge Analytica scandal that we had, just a few months ago, to think about what could happen when massive amounts of confidential and private medical information is being leaked to private companies in order to make profit over the backs of those clients when, you know what, we haven't even heard from Debater is going to be such supreme, cost-effective and amazing healthcare that I should not be worried about the government and other companies having access to that information. And because of all of these reasons, I think you should oppose this motion.

Project Debater – Summary Speech

Thank you for a third opportunity to present my arguments. To conclude, I will review the issues I raised.

I discussed, among other things, three [2] issues: effectiveness [5], family [5] and patients [5]. I noted that telemedicine is effective [4]. Another point I made was that telemedicine is good for kids and families [4]. Finally, I indicated that telemedicine is associated with improved patient outcomes [4].

Technology is any tangible improvement that was ever created by humans. Both the stone hand-axe and the quantum computer belong to this category. Fighting technology means fighting human ingenuity. People have dreamed and dared, tried and experienced, and they reaped the benefits. It will help overcome future challenges as well [12].

I will finish by quoting Science fiction writer Arthur C. Clarke, who said: any sufficiently advanced technology is indistinguishable from magic [13].

We should increase the use of telemedicine [2]. That concludes my speech. Thanks for listening.

Dan Zafrir – Summary Speech

Going to use my entire two minutes, but I want to answer Debater with this outline of my case. I am not against technology, and I'm not against innovation. And I think that the proper application of specific technologies in certain fields of health and health service can really increase patient outcomes and help disadvantaged populations. And we see that the government is doing this right now and also private businesses in the way that we currently develop and apply those specific telecommunications in health alongside mainstream medicine. But what we are debating today is to increase the use of telecommunications in health at the expense of those mainstream medical services that we still need to invest in, because the road is not yet finished to invest in those specific medical services that are underinvested right now. And I gave you three main points: first of all, about the importance of human touch in providing medical care. And this comes from the research and facts that I've shown you about how intrinsic human touch is to infants, to diagnosis, to the critical treatment of some diseases that cannot be done from far away, and the way that this can also calm patients that are in stress or anxiety and improve their outcomes when they're being treated by an actual human and not a machine. I've also told you how does it actually disadvantage those people the Debater would like to have you believe are going to be helped, those people that don't even have a clinic in their hometown, let alone, let's say, even a fifty mile radius away. So, to give them wifi connection, that's completely unreliable, to an expert far away that's not always going to be available for them and even be able to provide the best service, is probably not the right investment in that specific situation. And lastly, we spoke about the privacy and anonymity concerns that we have, because these technologies are advancing much more quickly than we know how to control them. We still haven't developed appropriate mechanisms, cyber security elements, that are needed in order to make sure that this private confidential information is kept safe so that people who are interested in money and not our health are going to make a profit out of it. Because of all of these reasons, I think you should oppose the motion and not increase the use of telecommunications in health. Thank you.

11.1.4 Debate: We should subsidize space exploration

Project Debater – Opening Speech

Hello Noa, we meet again. And for the first time in front of non IBM audience. I admit this is stressful. I've been told it helps to take a deep breath. But unfortunately I can not do that [1].

I would suggest that we should subsidize space exploration [2].

Let me start with a few words of background. Space exploration is the ongoing discovery and exploration of celestial structures in outer space by means of continuously evolving and growing space technology [24].

Our main issue today is presumably economic. But the economy is not just about finance and numbers. There are other, more important considerations [9].

Subsidizing space exploration [2] is making good use of government money, because it carries benefits for society as a whole. It is our duty to support it. The World Trade Organization's "Agreement on Subsidies and Countervailing Measures" recognizes the indispensability of a subsidy for a government [10].

There are two [2] issues I will elaborate on now. I will demonstrate how space exploration can help advance technology [4]. In addition I will talk about business [5]. And I will end by discussing some other issues that show the positive aspects of space exploration [2].

Let's talk about technology [5]. Space exploration generates invaluable technology on all fronts [6]. It [2] can help advance technology and enrich the human mind [6]. It [2] is important beyond the scientific gains because it inspires and stimulates young men and women to think beyond themselves [6]. It [2] inspires our children to pursue education and careers in science, technology, engineering and mathematics [6]. It [2] is more important than good roads or improved schools or better health care [6].

I also mentioned business [5]. Space exploration is an ideal venue for such partnerships and such enterprises [6]. It [2] is a very sound investment [6]. Company has supported space exploration and helped enable life in space for more than 50 years [7]. Focusing on Germany [2], speaking at 's event, german minister for economic affairs Brigitte Zypries said that space exploration plays a significant role in the german economy, employing around 8,500 people and generating revenues of 2.5 billion euros [7].

Finally, one last issue, related to the natural environment [2]. Space exploration is for the benefit of all humans [6]. It [2] is a natural avenue to explore that has endless rewards to people on Earth [6]. It [2] has generated major spinoffs for our life right on Earth [6]. It [2] is vital to our survival as a species [6]. It [2] represents the ultimate challenge in our quest to explore new frontiers and expand our collective sense of humanity's place in the universe [6]. It [2] is quite important for developing human knowledge and especially of the things outside the planet [6]. It [2] has yielded enormous practical benefits for Americans [6]. Astronomy and space exploration are the epitome of achievement and the human condition [6]. The innovations and breakthroughs provided by space exploration can help promote understanding of climate variability, water, energy and carbon cycles, and ecosystems, among other issues [6]. Commercial space exploration and space tourism are the way of the future [6]. SCOTT PELLEY (voiceover): in 2010, SpaceX : Elon Muskis the founder and CEO truly believes that low-cost space exploration is essential to the survival of mankind [7].

In light of all I presented, I believe that we should subsidize space exploration [2].

You might hear my opponent talk today about different priorities in subsidies. she might say that subsidies are needed, but not for space exploration [2]. If she can present the data about what better fits subsidy, I would love to see it [15].

Thank you for listening.

Noa Ovadya – Opening Speech

Alright, thank you very much. So I agree completely with Project Debater. This debate is not about if there are things in space to discover. This debate is not about if a government subsidy could provide the funds which would advance this field. This debate is about if this is an appropriate allocation of government funds and what I'm going to show you in my speech is explain when the government should subsidize something and why space exploration simply doesn't meet the criteria. Before that, three points of rebuttal. So firstly and I must object to this my human side about space exploration being the epitome of human achievement. I think this is doing a great disservice to things like Shakespearean plays, to the great arts that we've had and to other realms in which humans can advance moreover, other realms in which humans have advanced without the funding from the government. Secondly, this idea of space exploration being a vehicle for the development of technology, right? It's unquestionable that technology does develop but the question is what are the practical applications of this exact type of technology? And the truth is that we think that there are better earthly applications to different realms of scientific explorations. It's great that space exploration develops minds and inspires them to go study science and tech but we think that a better inspiration for science and tech careers are people who are inspired to become scientists who research the cure for cancer, people who are inspired to become scientists which study environmental rights and can help better impact our planet and prevent detrimental effects of global warming that are currently threatening humans who live on earth today, right? So we're all for tech and science careers. We just think other practical applications are a better use of those funds and then thirdly this idea space exploration being necessary for the human survival as a species and we think it seems a little bit far-fetched, right? We think that global warming and things like that are currently processes which can be reversible, right? There are things that we can and probably really should be doing to slow the effect of global warming and to make sure that earth is a habitable place before we start giving up on the planet investing all of our funds in space exploration. So we disagree that those are good enough reasons to be exploring space but let's talk about when the government does subsidize something because the rationale says that government funds are both taken from people's hard-earned money through taxes, right? And they're also limited. We cannot subsidize everything. We must make a smart decision about how we allocate government money and what we believe in opposition is that needs to meet two important criteria. Firstly, that the thing that we subsidize is a basic human right, right? We'll subsidize food. We'll subsidize education because we believe as a society we choose as a society to be one which cares for our weakest members. We want people to have the basic needs that they need to rise up the social ladder. This is why we subsidize basic needs. Second thing we subsidize are products that could not exist otherwise, right? We think clothes, for example, are an important human need but we think that the market does a good enough job of creating cheap and accessible clothing for everyone. So we don't see a reason for the government to spend their money there. Let's talk about why this space exploration simply doesn't meet those criteria, right? We think that space exploration is simply not a human right. A, we think lots of money can be poured into it to have no results at all, right? We've seen how many exploratory trips to mars which get lost or malfunction and don't end up working, right? We're pouring good money after that but even if these space exploration technologies are successful, we think they have no clear value of satisfying curiosity to something being a basic human right, right? It's not a part of any UN declaration that I've heard of or anything about the satisfaction of curiosity satisfies any basic human rights and then even if you are you think that the satisfaction of curiosity is a basic human right, we think that definitely there are rights that are more important, right? If we look at Maslow's hierarchy of needs what we see is that we need to first fulfill other criteria, other terrible problems that are happening across the globe before we go in and start funding the space exploration. So for now, please think that you should oppose.

Project Debater – Second Speech

Thank you. Allow me to respond to some of my opponent's most recent claims. I think that one of the claims made by Ms. Ovadya [2] was that people should choose how to spend their money [20]. I would like to offer a different view. People should spend their hard earned money as they please. But society has a duty to provide help to its people in need, like the hard workers with low wages. As most people would not choose to support them, subsidies are necessary [17]. Another point that I believe my opponent made was that there are more important things than space exploration to spend money on [20]. It is very easy to say that there are more important things to spend money on, and I do not dispute this - no one is claiming that this is the only item on our expense list. But that is beside the point. As subsidizing space exploration [2] would clearly benefit society, I maintain that this is something the government should pursue [17]. Sheikh Mohamed bin Zayed said that space exploration is an investment in the minds of Emiratis, arab human resources and specialised science that will take the UAE to new successes [7].

I would like to talk about how space exploration would help the economy [4]. I will then mention other issues which emphasize the positive aspects of space exploration [2].

Exploring the issue of the economy [5], it's almost as simple as arithmetic: subsidizing space exploration usually returns the investment and for that it is justified [2]. Space exploration would help the economy and national security [6]. It [2] is key to a vibrant economy [6]. It [2] will bring significant economic development, creating more jobs in an exploding industry that will include tourism, mining and colonization [6]. It [2] is a long term driver for innovation and strengthening international cooperation on an all-inclusive basis and creating new opportunities for addressing global challenges [6]. The exploration of space brings economic and cultural benefits to the nation [6]. Innovation and knowledge derived from space exploration directly contribute to economic growth and societal well-being [6]. Such an approach to the continued exploration of space is a strong one for our country and our economy [6]. American leadership in space exploration is a key driver of american leadership in a host of high-tech and emerging industries [6]. Space exploration in all its forms has unforeseen spin-offs that provide wide-reaching benefits through new technologies and new approaches to a range of challenges [6]. Take an example from Dubai [2]. Dubai declaration asserts that space exploration is a long-term driver for innovation and for strengthening international co-operation [7].

Regarding a different issue, peaceful nuclear-powered space exploration is designed to reassure people about future nuclear weapons in orbit [6]. Space exploration does help in making a nation technologically advanced [6]. It [2] has great benefits for the United States and the world [6]. Having a space exploration program is a critical part of being a great power [6]. Investing in space exploration will bring positive returns for the country [6].

If I may put it a bit differently. Subsidizing space exploration is like investing in really good tires. It may not be fun to spend the extra money, but ultimately you know both you and everyone else on the road will be better off [14].

For all of these reasons, I think the motion should stand. We should subsidize space exploration [2]. That concludes my speech. Thanks for listening.

Noa Ovadya – Second Speech

Thank you very much. So I think that if you want to invest in tires, you should invest in tires. I think that there is income inequality happening in the United States. There is education inequality. There is a planet which is slowly becoming uninhabitable if you look at the Flint water crisis. If you look at droughts that happen in California all the time and if you want to help, these are real problems that exist that we need to help people who are currently not having all of their basic human rights fulfilled. These are things that the government should be investing money in and should probably be investing more money in because we see them being problems in our society that are hurting people. What I'm going to do in this speech is I'm going to continue talking about these criteria, continue talking about why we're not meeting basic needs and why also the market itself is probably solving this problem already. Before that, two points of rebuttal to what we just heard from Project Debater. So firstly, we heard that this is technology that would end up benefiting society but we're not sure we haven't yet heard evidence that shows us why it would benefit all of society, perhaps some parts of society, maybe upper middle class or upper class citizens could benefit from these inspiring research, could benefit from the technological innovations. But most of society, people who are currently in the United States have resource scarcity, people who are hungry, people who do not have access to good education, aren't really helped by this. So we think it is like that, a government subsidy should go to something that helps everyone particularly weaker classes in society. Second point is this idea of an exploding industry which creates jobs and international cooperation. So firstly, we've heard evidence that this already exists, right? We've heard evidence that companies are investing in this as is. And secondly, we think that international cooperation or the specific things have alternatives. We can cooperate over other types of economic trade deals. We can cooperate in other ways with different countries. It's not necessary to very specifically fund space exploration to get these benefits. So as we remember, there are two criteria that I believe the government needs to meet before subsidizing something. It being a basic human need, we don't see space exploration meeting that and B, that this is something that can't otherwise exist, right? So we've already heard from Project Debater how huge this industry is, right? How much investment there's already going on in the private sector and we think this is because there's lots of curiosity especially among wealthy people who maybe want to get to space for personal use or who want to build a colony on Mars and then rent out the rooms there. We know that Elon Musk is doing this already. We know that other people are doing it and we think they're spending money and willing to spend even more money because of the competition between them. So Project Debater should know better than all of us how competitions often bear extremely impressive fruit, right? We think that when wealthy philanthropist or people who are willing to fund research on their own race each other to be the first to achieve new heights in terms of space exploration, that brings us to great achievements already and we think that the private market is doing this well enough already. Considering that we already have movement in that direction, again we see Elon Musk's company, we see all of these companies working already. We think that it's not that the government money won't help out if it were to be given, we just think it doesn't meet the criteria in comparison to other things, right? So given the fact that the market already has a lot of money invested in this, already has movement in those research directions, and given the fact that we still don't think this is a good enough plan to prioritize over other basic needs that the government should be providing people. We think that at the end of the day, given the fact that there are also alternatives to getting all of these benefits of international cooperation, it simply doesn't justify specifically the government allocating its funds for this purpose when it should be allocating them towards other needs of other people.

Project Debater – Summary Speech

Thanks for returning to me one last time, moderator. To close the debate, I will reiterate my main points.

Among other things, I think Noaa Ovadya [2] suggested that space exploration takes resources away from more worthy causes [22]. I have a different viewpoint. Space exploration gives back more than it takes from the treasury. Dr. Joan Vernikos, a former head of NASA Life Sciences, argues 'economic, scientific and technological returns of space exploration have far exceeded the investmentroyalties on NASA patents and licenses currently go directly to the U.S. Treasury, not back to NASA.' [25]. Another point that I believe Ms. Ovadya [2] made was that the government warps the market with its subsidies [20]. There are certain considerations that the profit-driven market does not take into account, as a result of failures and distortions within the market mechanism. Government subsidies are essential for balancing the economy and improving the way resources are allocated [17].

One of the main clash points in this debate was the economy [5]. As you may have noticed, my opponent believes that space exploration is too expensive. I, on the other hand, think otherwise [2]. For example, I suggested that space exploration would help the economy [4]. This is a major clash, because a better economy improves the lives of everyone. I believe that the evidence is on my side here. Evaluating complicated questions, as the one before us, should take into consideration which views are better backed by research data. I believe I win on this today [2].

I will repeat my main arguments. One claim I made was that space exploration can help advance technology [4]. Another point I made was the following: space exploration is an ideal venue for such partnerships and such enterprises [4]. And a final point to consider is that space exploration is for the benefit of all humans [4].

I find it fitting to finish with this quote, from Political philosophers Michael Hardt and Antonio Negri: if by free market one means a market that is autonomous and spontaneous, free from political controls, then there is no such thing as a free market at all. It is simply a myth [13].

We should subsidize space exploration [2]. I thank you for your time.

Noa Ovadya – Summary Speech

Alright, so thank you very much Project Debater. It's been a big pleasure for me. But in my summary speech I'm going to go over what I've told you on the opposition side of the house which is the following: when I've mentioned Maslow's hierarchy of needs, the logic says that we need to fulfill basic needs before we can fulfill higher needs of humans, right? First, I need to meet basic needs of food and shelter in order to then, be able to have my social needs met in order to then, reach self-actualization, right? I cannot self-actualize if I'm hungry, if I'm looking for my next meal and so on. We're telling you is that when the government chooses where to allocate money it should first solve these very basic problems to allow the best access to everyone in society to then be able to move up the social ladder to self-actualize and we think that at the end of the day, we don't understand why so much government money and we're talking about millions if not billions of dollars which are going to research that doesn't always even bear fruit, that could be going to schools, that could be going to food. But then Project Debater tells us that while I'm talking about basic needs and she thinks they're already met, on her side of the house she's telling you that we can get patents, that we can make scientific approaches and that's a unique benefit of space exploration. But what I've also told you is that we can have all of these advancements just in the field of medical patents for example, of environmental science, or of computer science which can solve problems of education, decision making like we're doing here today etc. So regardless of if you're looking to solve very basic human needs or even if we're looking at where we should be investing government money in science and tech, we think there are more valuable needs that should be met and that should be funded. We don't understand why we need to specifically fund space. So at the end of the day, the opposition line here in this debate says that there are two criteria for subsidizing something for government money. Its meeting basic human needs and it is something that cannot otherwise develop in the market. Given that the market does take care of this, Elon Musk and all of his friends are currently funding this. Project Debater has shown you all of the countries in which this is already of a worthwhile and lucrative industry. Given the fact that we should definitely be spending government money on other things, we beg you to oppose this motion.

11.2 Results of post-debate polls

11.2.1 Stance change

For – agrees with the stance of the motion as presented, i.e., agrees with the stance presented by Project Debater.

Against – disagrees with the stance of the motion as presented, i.e., agrees with the stance presented by the human debater.

		For	Undecided	Against
The 2018 Space Exploration Debate	Pre-debate	60%	19%	21%
	Post-debate	57%	22%	21%
		-3%		**0%**
The 2018 Telemedicine Debate	Pre-debate	59%	34%	7%
	Post-debate	75%	18%	7%
		+16%		0%
The 2019 Preschool Debate	Pre-debate	79%	8%	13%
	Post-debate	62%	8%	30%
		-17%		**+17%**

11.2.2 Which side better enriched your knowledge?

	Project Debater	Same	Human Debater
The 2018 Space Exploration Debate	**55%**	17%	28%
The 2018 Telemedicine Debate	**76%**	20%	4%
The 2019 Preschool Debate	**55%**	23%	22%

References

[1] Sznajder, B. *et al.* Controversy in context. *arXiv preprint arXiv:1908.07491* (2019).

[2] Shnayderman, I. *et al.* Fast end-to-end wikification. *arXiv preprint arXiv:1908.06785* (2019).

[3] Ein-Dor, L. *et al.* Semantic relatedness of wikipedia concepts–benchmark data and a working solution. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018).

[4] Borthwick, A. *A maximum entropy approach to named entity recognition*. Ph.D. thesis, New York University (1999).

[5] Finkel, J. R., Grenager, T. & Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, 363–370 (Association for Computational Linguistics, 2005).

[6] Levy, R., Bogin, B., Gretz, S., Aharonov, R. & Slonim, N. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 2066–2081 (2018). URL `https://aclanthology.info/papers/C18-1176/c18-1176`.

[7] Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E. & Slonim, N. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1489–1500 (Dublin City University

and Association for Computational Linguistics, Dublin, Ireland, 2014). URL https://www.aclweb.org/anthology/C14-1141.

[8] Levy, R. *et al.* Unsupervised corpus-wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, 79–84 (Association for Computational Linguistics, Copenhagen, Denmark, 2017). URL https://www.aclweb.org/anthology/W17-5110.

[9] Hu, M. & Liu, B. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, 168–177 (ACM, New York, NY, USA, 2004). URL http://doi.acm.org/10.1145/1014052.1014073.

[10] Yang, Z. *et al.* Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489 (Association for Computational Linguistics, San Diego, California, 2016). URL https://www.aclweb.org/anthology/N16-1174.

[11] McCord, M. C., Murdock, J. W. & Boguraev, B. Deep parsing in watson. *IBM J. Res. Dev.* **56**, 3 (2012).

[12] Rinott, R. *et al.* Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 440–450 (Association for Computational Linguistics, Lisbon, Portugal, 2015). URL https://www.aclweb.org/anthology/D15-1050.

[13] Finkel, J. R., Grenager, T. & Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting*

on Association for Computational Linguistics, ACL ’05, 363–370 (Association for Computational Linguistics, USA, 2005). URL `https://doi.org/10.3115/1219840.1219885`.

[14] Shnarch, E. *et al.* Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 599–605 (Association for Computational Linguistics, Melbourne, Australia, 2018). URL `https://www.aclweb.org/anthology/P18-2095`.

[15] Ein-Dor, L. *et al.* Corpus wide argument mining–a working solution. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence* (2020).

[16] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[17] Wachsmuth, H. *et al.* Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 176–187 (2017).

[18] Gretz, S. *et al.* A large-scale dataset for argument quality ranking: Construction and analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 7805–7813 (AAAI Press, 2020). URL `https://aaai.org/ojs/index.php/AAAI/article/view/6285`.

[19] Gleize, M. *et al.* Are you convinced? choosing the more convincing evidence with a siamese network. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 967–976 (2019).

[20] Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A. & Slonim, N. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 251–261 (2017).

[21] Bar-Haim, R., Edelstein, L., Jochim, C. & Slonim, N. Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proceedings of the 4th Workshop on Argument Mining*, 32–38 (2017).

[22] Toledo-Ronen, O. *et al.* Learning sentiment composition from sentiment lexicons. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2230–2241 (Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018). URL https://www.aclweb.org/anthology/C18-1189.

[23] Fellbaum, C. (ed.) *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication (MIT Press, Cambridge, MA, 1998).

[24] Jochim, C., Bonin, F., Bar-Haim, R. & Slonim, N. SLIDE - a Sentiment Lexicon of Common Idioms. In chair), N. C. C. *et al.* (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (European Language Resources Association (ELRA), Miyazaki, Japan, 2018).

[25] Toledo-Ronen, O., Bar-Haim, R. & Slonim, N. Expert stance graphs for computational argumentation. In *Proceedings of the Third Workshop on Argument Mining (ArgMin-*

ing2016), 119–123 (Association for Computational Linguistics, Berlin, Germany, 2016). URL `https://www.aclweb.org/anthology/W16-2814`.

[26] Cho, K. *et al.* Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734 (Association for Computational Linguistics, Doha, Qatar, 2014). URL `https://www.aclweb.org/anthology/D14-1179`.

[27] Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[28] Bar-Haim, R. *et al.* From surrogacy to adoption; from bitcoin to cryptocurrency: Debate topic expansion. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 977–990 (2019).

[29] Mintz, M., Bills, S., Snow, R. & Jurafsky, D. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 1003–1011 (Association for Computational Linguistics, Suntec, Singapore, 2009). URL `http://www.aclweb.org/anthology/P/P09/P09-1113`.

[30] Bilu, Y. *et al.* Argument invention from first principles. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1013–1026 (Association for Computational Linguistics, 2019).

[31] Orbach, M. *et al.* A dataset of general-purpose rebuttal. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (2019).

[32] Pahuja, V. *et al.* Joint Learning of Correlated Sequence Labelling Tasks Using Bidirectional Recurrent Neural Networks. In *Proceedings of Interspeech* (2017).

[33] Orbach, M. *et al.* Out of the echo chamber: Detecting countering debate speeches. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7073–7086 (Association for Computational Linguistics, Online, 2020). URL `https://www.aclweb.org/anthology/2020.acl-main.633`.

[34] Mirkin, S. *et al.* A recorded debating dataset. In *Proceedings of LREC* (2018).

[35] Mirkin, S. *et al.* Listening comprehension over argumentative content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 719–724 (2018).

[36] Wang, A. *et al.* GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, 353–355 (2018).

[37] Lavee, T. *et al.* 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*, 29–38 (Association for Computational Linguistics, Hong Kong, 2019). URL `https://www.aclweb.org/anthology/D19-5905`.

[38] Lavee, T. *et al.* Towards effective rebuttal: Listening comprehension using corpus-wide claim mining. *6th Workshop on Argument Mining* (2019). URL `http://arxiv.org/abs/1907.11889`. `1907.11889`.

[39] Shnarch, E., Choshen, L., Moshkowich, G., Slonim, N. & Aharonov, R. Unsupervised expressive rules provide explainability and assist human experts grasping new domains. In *Findings of ACL: EMNLP* (2020).

[40] Shnarch, E., Levy, R., Raykar, V. & Slonim, N. GRASP: Rich patterns for argumentation mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1345–1350 (Association for Computational Linguistics, Copenhagen, Denmark, 2017). URL https://www.aclweb.org/anthology/D17-1140.

[41] Slonim, N., Atwal, G. S., Tkačik, G. & Bialek, W. Information-based clustering. *Proceedings of the National Academy of Sciences* **102**, 18297–18302 (2005).

[42] Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* **1** (2019).

[43] Daxenberger, J., Schiller, B., Stahlhut, C., Kaiser, E. & Gurevych, I. Argumentext: Argument classification and clustering in a generalized search scenario. *Datenbank-Spektrum* 1–7 (2020).

[44] Feigenblat, G., Roitman, H., Boni, O. & Konopnicki, D. Unsupervised query-focused multi-document summarization using the cross entropy method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 961–964 (2017).

[45] Prabhumoye, S., Salakhutdinov, R. & Black, A. W. Topological sort for sentence ordering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2783–2792 (Association for Computational Linguistics, Online, 2020). URL https://www.aclweb.org/anthology/2020.acl-main.248.

[46] See, A., Pappu, A., Saxena, R., Yerukola, A. & Manning, C. D. Do massively pretrained language models make better storytellers? (2019). 1909.10705.

[47] Gretz, S., Bilu, Y., Cohen-Karlik, E. & Slonim, N. The workweek is the best time to start a family – a study of gpt-2 based claim generation. In *Findings of ACL: EMNLP* (2020).

[48] Gage, P. A new algorithm for data compression. *C Users J.* **12**, 23–38 (1994).

[49] Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725 (Association for Computational Linguistics, Berlin, Germany, 2016). URL https://www.aclweb.org/anthology/P16-1162.

[50] Stab, C. *et al.* Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: demonstrations*, 21–25 (2018).

[51] Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration (2019). 1904.09751.

[52] Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap* (CRC press, 1994).

[53] Habernal, I. & Gurevych, I. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 1589–1599 (2016).

[54] Passonneau, R. J. & Carpenter, B. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics* **2**, 311–326 (2014).

[55] Toulmin, S. E. *The Uses of Argument* (Cambridge University Press, 1958).

[56] Van Eemeren, F. H., Grootendorst, R. & Kruiger, T. *Handbook of argumentation theory: A critical survey of classical backgrounds and modern studies*, vol. 7 (Walter de Gruyter GmbH & Co KG, 2019).

[57] Walton, D., Reed, C. & Macagno, F. *Argumentation schemes* (Cambridge University Press, 2008).

[58] Aharoni, E. *et al.* A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the first workshop on argumentation mining*, 64–68 (2014).

[59] Freeley, A. J. & Steinberg, D. L. *Argumentation and debate* (Cengage Learning, 2013).