Listening Comprehension over Argumentative Content

Shachar Mirkin[*1], **Guy Moshkowich**[*2], **Matan Orbach**[*2], **Lili Kotlerman**[2], **Yoav Kantor**[2],
Tamar Lavee[3], **Michal Jacovi**[2], **Yonatan Bilu**[2], **Ranit Aharonov**[2] **and Noam Slonim**[2]
[1]Digimind, [2]IBM Research, [3]IBM Watson

Abstract

This paper presents a task for machine listening comprehension in the argumentation domain and a corresponding dataset in English. We recorded 200 spontaneous speeches arguing for or against 50 controversial topics. For each speech, we formulated a question, aimed at confirming or rejecting the occurrence of potential arguments in the speech. Labels were collected by listening to the speech and marking which arguments were mentioned by the speaker. We applied baseline methods addressing the task, to be used as a benchmark for future work over this dataset. All data used in this work is freely available for research.

1 Introduction

Machine reading comprehension (MRC) is the NLP task equivalent to reading comprehension tests that assess the understanding of written texts by humans. MRC is usually realized as a question answering (QA) task through multiple-choice questions or as a cloze test (Richardson et al., 2013; Hermann et al., 2015; Hill et al., 2015). With the abundance of multimedia content nowadays, this line of work has been extended to speech, by applying QA methods to speech transcripts, i.e. the output of automatic speech recognition (ASR). In such works, the task is consequently referred to as ‘spoken question answering’ (Li et al., 2018), ‘question answering over speech transcripts’ (Turmo et al., 2007; Lamel et al., 2008) or *machine listening comprehension* (MLC) (Chung and Glass, 2018).

We continue this line of work, and present a listening comprehension task and associated benchmark data over argumentative content. In the argumentation domain, such as political debates, people are often exposed directly to the audio (or the video), without access to a written version. Human comprehension is then done in real-time through listening. We simulate this scenario in our dataset. Namely, annotation is carried out by *listening* to debate speeches rather than by reading transcripts as done in prior work. The auditory modality is richer than written text in terms of the signal available to listeners, e.g., prosody. Similarly, machine comprehension can make use of the extra-lexical signal. The dataset we construct and release enables utilizing such signals, as done for instance in (Lippi and Torroni, 2016) for detecting claims in debates.

Most often, in both reading and listening comprehension tasks, the answer is explicitly mentioned in the text; frequently, the answer is even an actual segment of the text, as in SQuAD (Rajpurkar et al., 2016), one of the most popular MRC datasets. Conversely, in argumentation, presuppositions are fundamental (Habernal et al., 2018), inferences are more subtle and the answer may rely on common knowledge. Going beyond the factoid level, Tseng et al. (2016) presented a listening comprehension task over TOEFL listening tests.[1] In comparison, our data consists of spontaneous speech and is not adapted for non-native speakers.

We use data from iDebate[2], a high-quality, curated database of controversial topics – referred to as “motions”, as in formal parliamentary proposals – with a list of arguments for and against each motion. We selected 50 motions, and recorded experienced debaters making four speeches for each of them (two for and two against the motion). We then asked annotators to listen to a speech and presented them with a list of arguments that were proposed independently in iDebate for the motion. The annotators had to mark which of these argu-

[*] This work was done at IBM within Project Debater; the first 3 authors equally contributed to this work.

[1] https://www.ets.org/toefl

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

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 719–724
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

ments were mentioned in the speech (see Section 2 for further details). Example 1 shows one such argument alongside a speech snippet, demonstrating the unique nature of this domain and data. Specifically, the argument against the motion is implied from the speech, but is not explicitly mentioned in it.

Example 1 (Positively labeled argument)
Motion: *We should introduce goal line technology [in sports]*
Argument: *Controversy and debate are a part of the game*
Speech stance: *opposing ("con")*
... people also like it to some extent when officials make mistakes, because it adds to some of the like drama, the, oh, what if this happened? ... And I think that one of the biggest things that fans enjoy bonding over is when refs make mistakes that are blatantly wrong.

iDebate was chosen since its arguments are good candidates to construct comprehension questions: more than half of the assessed arguments are mentioned in our recorded speeches, and the large majority of the speeches contain at least one of the arguments suggested by iDebate. Furthermore, each argument in iDebate is coupled with a counter-argument. Those, in turn, may be used to rebut each argument that was detected through MLC, suggesting intriguing future directions to explore the released data. In a task related to ours, Boltužić and Šnajder (2014) have searched iDebate arguments in user argumentative comments. Their work, though, consisted of only two motions and included written, rather than audio data.

We release our annotated data and the results of baseline methods applied to it as a benchmark dataset for the MLC task. The dataset includes 200 speeches for 50 motions, in English. For each speech we include the following: (i) the audio version of the speech; (ii) manual and automatic transcripts; (iii) a labeled listening comprehension question, consisting of a set of arguments from iDebate that potentially appear in the speech.

The main contributions of this work are: (i) proposal of the new task of listening comprehension over argumentative content, a domain very different from those previously used for reading or listening comprehension tasks; (ii) a comprehensive and rich labeled dataset of 200 speeches covering 50 topics, transcribed both automatically and manually, and labeled for the listening comprehension task; (iii) establishment of baselines over the dataset.

All the recordings, their transcripts and labels are available for research at `https://www.research.ibm.com/haifa/dept/vst/debating_data.shtml`.

2 Data

iDebate iDebate contains a list of controversial topics, phrased as parliamentary motions. Each motion is associated with arguments (referred to as "points") supporting or contesting it. Each argument may be comprised of several sentences and is briefly summarized in a *title*.

Selecting iDebate motions At the time of this research, iDebate contained 684 motions.[3] We selected 50 clearly-defined motions, and simplified their phrasing when necessary. For example, we rephrased "This House believes that cannabis should be legalised" to "We should legalize cannabis".

Producing debate speeches We recorded argumentative speeches for each motion. First, two speeches supporting each motion were recorded by two experienced debaters. In doing so, we followed the process described in (Mirkin et al., 2018), where a speaker is presented with a motion and its description and is instructed to record a few minutes speech that supports it, with 10 minutes to prepare, but without checking any online materials. Given a speech supporting the motion, we asked another debater to listen to it and then record a speech rebutting it, and in consequence – opposing the motion. These response speeches are of different nature than the initial speeches beyond the opposite stance, as they often contain references and rebuttal to arguments mentioned in this initial speech.

Through this process, we produced in total 200 speeches recorded by 14 different speakers, equally distributed between the motions, and between the "pro" and "con" stances. Each speech was transcribed automatically using Watson ASR.[4] The transcripts were split automatically into sentences using a bi-directional LSTM that was trained on spoken language corpora (Pahuja et al., 2017). For completeness of the dataset,

[3] We accessed iDebate on Jan. 28, 2018.

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

for each ASR transcript we include a manually-corrected "reference" human transcript, including manually added punctuation. Based on the human transcripts, we computed the word error rate (WER) of the ASR transcripts: 8.03% on average. For comparison, the WER of the ASR transcripts in (Li et al., 2018) is 22.73%.

Labeling Given the recorded speeches, we carried out a labeling task employing experienced annotators, all of whom are highly proficient English-speakers. Given a motion and a speech, the annotators were instructed to listen to it once, preferably without pausing, and select which ones of iDebate argument titles were mentioned in it, or *None* if none of them was. Specifically, they were instructed to answer positively if it would be correct to say that "the speaker argued that arg ..." where arg is the argument's title. A single question contained all the iDebate titles for the motion, which have the suitable stance for the given speech. Each of the 200 questions was answered by five annotators.

On average, the labeled data contains 4.4 argument titles per speech, where a title contains 10.5 words and an argument text includes 6 sentences and 156 words.

Labeling results In 173 (86.5%) of the speeches, at least one iDebate argument was found, and 248 (~56%) of the iDebate arguments were labeled as positive at least once.

In order to analyze agreement between annotators, we transformed each multiple-choice question to a set of binary questions containing a speech and a single argument title. This amounted to 878 annotated speech–argument pairs, of which 354 (40.3%) are labeled as positive (i.e. an average of ~1.8 positive arguments per speech). The average pairwise Cohen's Kappa (Cohen, 1960) score over the labels is 0.52 (0.55 for supporting and 0.50 for opposing speeches). The Fleiss' Kappa (Fleiss, 1971) is 0.52. Noteworthy, 78.5% of the labels were of high confidence: four or five annotators agreed on the label. Figure 1 shows the distribution of positive answers over the binary questions.

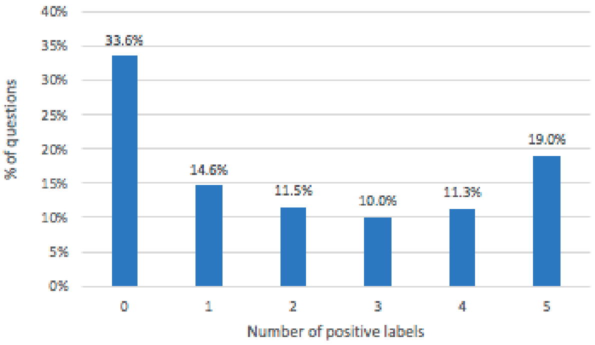

Figure 1: The percentage of binary questions labeled as positive by 0 to 5 annotators.

We analyzed a sample of arguments–speech pairs that had low agreement between annotators, i.e., those that have two or three positive labels. One reason for disagreement that we identified concerns argument titles that contain two claims, of which only one is argued by the speaker. An example is the title "Gambling is associated with other forms of addiction and harmful behaviour" concerning the motion "We should ban gambling". In a speech about this motion, only "addiction" was mentioned but not "harmful behaviour", resulting with two annotators accepting it and three rejecting it. Another possible source of disagreement is argument titles that are semantically similar, but not identical to the arguments presented in the speech. For instance, two of five annotators accepted the argument "on-line gambling affects families" when the speaker argued on the effects of gambling on families, but did not mention on-line gambling specifically.

Listening vs. reading To corroborate the reliability of our labeling through listening, we repeated the task for 40 randomly sampled speeches, replacing only the audio with manual human transcripts of the speeches.[5] The average pairwise Cohen's Kappa score over the labels is 0.59 and the Fleiss' Kappa is 0.60. While these may indicate that the reading task is somewhat easier (e.g. because the annotator can read the text multiple times), it was encouraging to discover that audio-based labeling achieves similar results to text-based labeling: we compared the labels obtained via reading and via listening and found that 87% of them were identical. Labeling by listening is closer to the task of listening comprehension than labeling via reading. Another advantage is that it removes the need to manually transcribe the speeches (in our experience, ASR transcripts are not ideal for labeling). As mentioned, in our

[5] We use manual and not ASR transcripts for this analysis, under the assumption that when listening, the annotators are also receiving error-free content.

released dataset we do provide human transcripts for all speeches.

Analysis of potential bias Recent works showed undesired artifacts in natural language inference datasets, namely that in some datasets for inferring relations between two texts, inference can in fact be done by only considering one of them (Schwartz et al., 2017; Gururangan et al., 2018; Tsuchiya, 2018; Poliak et al., 2018). To explore this issue in our dataset, we assessed the correlation between several features of the argument title and the label. Specifically, we computed the Spearman's correlation (Spearman, 1904) between the label and the title's length, occurrences of named entities or negation words in the title, and the correlation between the labels and the titles' 100 most frequent words, stopwords excluded. This resulted in low correlation coefficients, summarized in Table 1. This preliminary analysis suggests that naive features extracted from the title are not sufficient for predicting its label.[6]

Feature	**Spearman ρ**
Title length	-0.07
Named entity	-0.08
Negation	-0.03
Frequent words	0.14

Table 1: Spearman correlation between labels and features of argument titles. For the frequent words, the figure shown is the maximum (absolute) correlation found between positive labels and words from the titles. The words which yielded this correlation were “women” and “environment”.

3 Evaluation

Next, we establish baselines for our annotated dataset. All baselines are based on simple unsupervised text similarity methods for selecting which arguments were mentioned in a debate speech. Strictly, this is an entailment task rather than a similarity task, and similarity serves here as an approximation. Below we describe the establishment of the baselines.

Evaluation configurations We considered two ways of representing a speech and two ways of representing an argument, to a total of four different experimental configurations. A speech can be represented as a single text or as the set of its sentences. When a speech is represented by a set of sentences, we considered the maximum similarity score obtained by a sentence in the speech. For the speech text we used the ASR transcripts of the audio speeches, with the automatic split into sentences, as described in Section 2. The sentence-based configurations, therefore, depend on the specific automatic splitting of the texts into sentences. An argument can be represented using only its title or by its extended text. The validity of matching an argument's text against the speech while labels refer to titles stems from entailment transitivity: under the observation that in our data an argument text typically entails its title, if an argument text is mentioned in the speech (i.e. is entailed by it), its title is also entailed by the speech (that is, given that typically $arg \Rightarrow title$, if $speech \Rightarrow arg$ then $speech \Rightarrow title$.)

Evaluation metric The performance of each method is calculated as the average accuracy over speeches in the test set, where the accuracy of a speech is the ratio of correct answers out of the number of choices presented for it. This ensures that each speech contributes equally to the overall accuracy regardless of the number of potential arguments associated with it. Since there is an equal number of speeches per motion, this is also the average accuracy over motions.[7]

Development and test sets We randomly split the data into development and test sets (*dev* and *test* below), such that 60% of the motions (30 motions, 120 speeches, 60 of each of supporting and opposing speeches) are in *dev* and 40% (80 speeches) are in *test*. For each method, we select a threshold that maximizes accuracy over *dev*, and apply it to *test*. In other words, an argument is considered to be mentioned in the speech if its similarity score is above the threshold.

[6] We thank the anonymous reviewers for helping us improve this analysis.

[7] We chose accuracy over precision and recall since we wished to give an equal weight to each question; therefore, a micro-average score – where we consider each argument-speech pair as an item in the calculation – was unsuitable. With a small number of pairs per question, one often encounters anomalies: when there are no positive labels, recall is undefined, and when no positive prediction is given for a question, precision is undefined.

3.1 Assessed methods

***All-yes* baseline** As a reference point, we compute the accuracy obtained when all arguments are predicted to be mentioned in the speech, resulting in 39.8% accuracy.

word2vec (*w2v*) We create a w2v (Mikolov et al., 2013) vector representation for each text, removing stopwords; each word is represented by a 200-dimensional word embedding learned over Wikipedia. A *tfidf*-weighted average of the word embeddings represents each text, where *idf* values are counted when considering each Wikipedia sentence as a document. Given a pair of texts, their score is the cosine similarity between their vector representations.

skip-thought (ST) Kiros et al. (2015) presented a general sentence encoder, that has been applied successfully to a variety of tasks such as semantic relatedness and paraphrase detection, often obtaining state of the art results. We use its available implementation to encode the texts as vectors, and compute the cosine similarity between them.

3.2 Results

Table 2 shows the accuracy of all *w2v* configurations. Representing an argument using its more verbose several-sentences-long content outperforms using its short single-sentence title. On the speech side, considering each sentence separately is preferable to using the entire speech. We compared the results of the best *w2v*-based configuration (*arg-sentence*), to the performance of the *skip-thought* auto-encoder. In this setting, encoding individual speech sentences and an argument, the accuracy of *skip-thought* was 60.2%.

Method	Accuracy (%)
all-yes	39.8
w2v title-speech	49.8
w2v arg-speech	57.6
w2v title-sentence	55.8
w2v arg-sentence	64.6
ST arg-sentence	60.2

Table 2: Accuracy results over the test set ASR transcripts, for *w2v* and *skip-thought* (ST).

The highest scoring method, *w2v arg-sentence*, reaches, then, a rather modest accuracy of 64.6%. One weakness of this method, revealed through analysis of its false positive predictions, is its tendency to prefer longer sentences. It is nevertheless substantially superior to the trivial *all-yes* baseline, as well as its *all-no* counterpart.

As explained, we chose accuracy as the main metric for this benchmark as it enables computing macro-average scores over speeches with a variable number of arguments. For reference, the micro-average precision and recall scores of the *w2v arg-sentence* are 57.1% and 43% respectively, with an F_1 score of 49.1%. Optimizing for this metric enables controlling the trade-off between precision and recall with a threshold, depending on the end-application needs.

In sum, we have set a baseline for this task by computing similarity between averaged word embeddings vectors. This simple method can be used as a starting point for future works on this dataset.

4 Conclusions

Machine listening comprehension is a challenging task, whose complexity stems, among other things, from the difficulty to handle spoken language and from errors due to automatic transcription. The argumentation domain, often with complex and elaborate reasoning, relying on presuppositions and world knowledge, adds another dimension to this complexity. In this work, we suggest a task and a corresponding benchmark dataset to assess comprehension in this domain. We focused on the task of confirming the occurrence of arguments in a speech, which – as shown in this work – can be handled to some degree with standard textual inference methods. Other types of questions can be formulated over this data in following work. We release a rich dataset, accompanied with benchmarks, that can drive various studies in listening comprehension and argumentation mining.

5 Acknowledgments

We are thankful to the debaters and annotators who took part in the creation of this dataset.

References

F. Boltužić and J. Šnajder. 2014. Back up your Stance: Recognizing Arguments in Online Discussions. In *Proceedings of the First Workshop on Argumentation Mining, ACL.*

Yu-An Chung and James R. Glass. 2018. Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech. *CoRR*, abs/1803.08976.

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

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76(5):378–382.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. *Proceedings of NAACL*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. In *Proceedings of NAACL*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of NIPS*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *CoRR*, abs/1511.02301.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. *NIPS*.

Lori Lamel, Sophie Rosset, Christelle Ayache, Djamel Mostefa, Jordi Turmo, and Pere Comas. 2008. Question Answering on Speech Transcriptions: the QAST evaluation in CLEF. In *Proceedings of LREC*.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hungyi Lee. 2018. Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension. In *Proceedings of Interspeech*.

Marco Lippi and Paolo Torroni. 2016. Argument Mining from Speech: Detecting Claims in Political Debates. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*.

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

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

Vardaan Pahuja, Anirban Laha, Shachar Mirkin, Vikas Raykar, Lili Kotlerman, and Guy Lev. 2017. Joint Learning of Correlated Sequence Labelling Tasks Using Bidirectional Recurrent Neural Networks.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of EMNLP*.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. *Proceedings of CoNLL*.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards Machine Comprehension of Spoken Content: Initial TOEFL Listening Comprehension Test by Machine. In *Proceedings of Interspeech*.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of LREC*.

Jordi Turmo, Pere Comas, Christelle Ayache, Djamel Mostefa, Sophie Rosset, and Lori Lamel. 2007. Overview of qast 2007. In *Proceedings of CLEF*.