Crowd-sourcing annotation of complex NLU tasks: A case study of argumentative content annotation

Tamar Lavee, Lili Kotlerman,[*] Matan Orbach, Yonatan Bilu,
Michal Jacovi, Ranit Aharonov and Noam Slonim
IBM Research

Abstract

Recent advancements in machine reading and listening comprehension involve the annotation of long texts. Such tasks are typically time consuming, making crowd-annotations an attractive solution, yet their complexity often makes such a solution unfeasible. In particular, a major concern is that crowd annotators may be tempted to skim through long texts, and answer questions without reading thoroughly. We present a case study of adapting this type of task to the crowd. The task is to identify claims in a several minute long debate speech. We show that sentence-by-sentence annotation does not scale and that labeling only a subset of sentences is insufficient. Instead, we propose a scheme for effectively performing the full, complex task with crowd annotators, allowing the collection of large scale annotated datasets. We believe that the encountered challenges and pitfalls, as well as lessons learned, are relevant in general when collecting data for large scale natural language understanding (NLU) tasks.

1 Introduction

The availability and scale of crowdsourcing platforms today has enabled the collection of large scale labeled datasets (Negri et al., 2011; Sabou et al., 2014; Rajpurkar et al., 2016, 2018; Choi et al., 2018). These datasets facilitate the use of advanced machine learning methods, which leverage such vast volumes of labeled data to achieve state-of-the-art performance on various tasks. Crowd annotation tasks are typically simple, short, and easy to explain, making them well-suited to the typically untrained temporary workforce. Some examples include named entity recognition (Finin et al., 2010), textual entailment (Mehdad et al., 2010) or generating facts from text (Wang and Callison-Burch, 2010). Complex tasks are typically broken into smaller, simpler chunks to suit these requirements (Wang et al., 2013). For example, Zeichner et al. (2012) break up their evaluation of inference rules into three simpler sub-tasks, and Scholman and Demberg (2017) simplify their discourse relation annotation task by casting it as a selection of a connecting phrase from a predefined list. Indeed, GLUE (Wang et al., 2018), a popular benchmark for NLU tasks, focuses only on annotations of single sentences or pairs of sentences, which tend to be simpler than those required in longer texts. However, task decomposition is not always feasible. As we discuss below, while a relevant decomposition scheme can be defined for our task, it does not allow performing the task in an effective and comprehensive way.

We describe the adaptation of a complex labeling task to the crowd: identifying claims in spoken argumentative content (for an example, see Figure 1). This work extends our previous study, in which annotation was performed by experts (Mirkin et al., 2018).

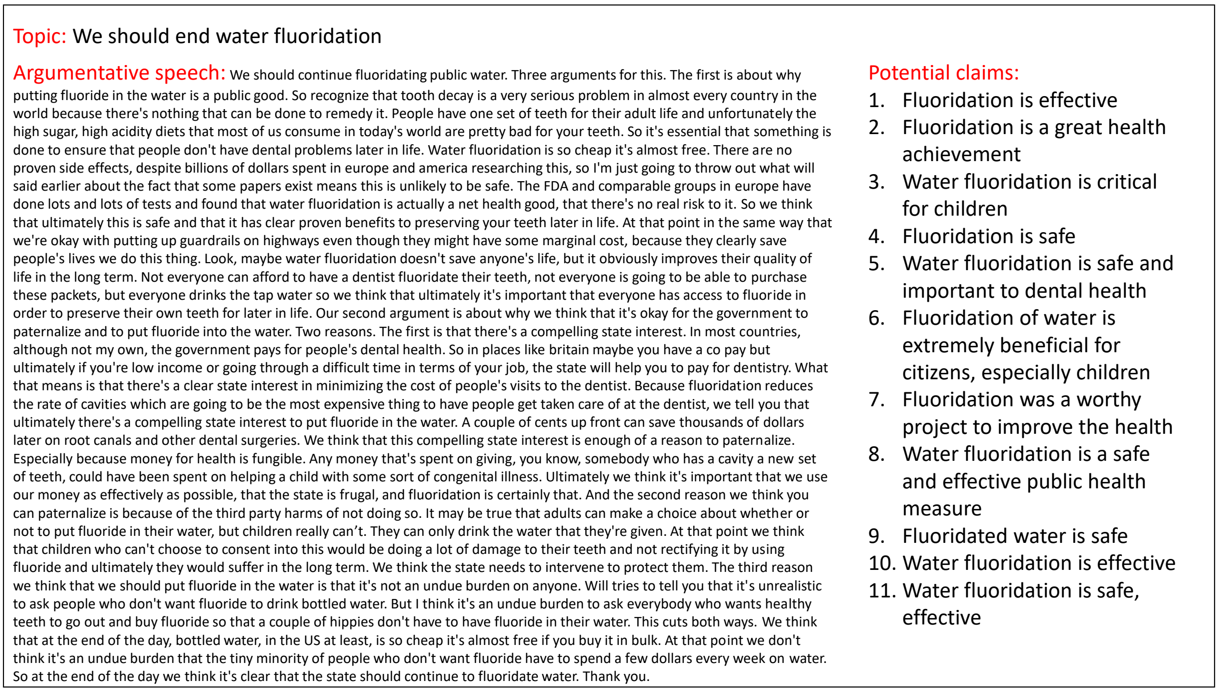
Obtaining such labeled data facilitates the development of language understanding systems which listen to speeches and identify claims therein. This, in turn, can serve as the basic building block for generating arguments rebutting these claims, or summarizing an argumentative text into the main claims made therein. Indeed, this annotation was made in the context of Project Debater, a system that can hold a debate with humans[1], where rebuttal was based on Argument Mining (Lavee et al., 2019) and general-purpose claims (Orbach et al., 2019).

At first glance, simplifying such a task could seem straightforward. By segmenting speeches

[*] Current affiliation: Intuition Robotics

[1] Demonstrated at Think 2019; https://www.youtube.com/watch?v=m3u-1yttrVw

Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP, pages 29–38
Hong Kong, China, November 3, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17



Topic: We should end water fluoridation

Argumentative speech: We should continue fluoridating public water. Three arguments for this. The first is about why putting fluoride in the water is a public good. So recognize that tooth decay is a very serious problem in almost every country in the world because there's nothing that can be done to remedy it. People have one set of teeth for their adult life and unfortunately the high sugar, high acidity diets that most of us consume in today's world are pretty bad for your teeth. So it's essential that something is done to ensure that people don't have dental problems later in life. Water fluoridation is so cheap it's almost free. There are no proven side effects, despite billions of dollars spent in europe and america researching this, so I'm just going to throw out what will said earlier about the fact that some papers exist means this is unlikely to be safe. The FDA and comparable groups in europe have done lots and lots of tests and found that water fluoridation is actually a net health good, that there's no real risk to it. So we think that ultimately this is safe and that it has clear proven benefits to preserving your teeth later in life. At that point in the same way that we're okay with putting up guardrails on highways even though they might have some marginal cost, because they clearly save people's lives we do this thing. Look, maybe water fluoridation doesn't save anyone's life, but it obviously improves their quality of life in the long term. Not everyone can afford to have a dentist fluoridate their teeth, not everyone is going to be able to purchase these packets, but everyone drinks the tap water so we think that ultimately it's important that everyone has access to fluoride in order to preserve their own teeth for later in life. Our second argument is about why we think that it's okay for the government to paternalize and to put fluoride into the water. Two reasons. The first is that there's a compelling state interest. In most countries, although not my own, the government pays for people's dental health. So in places like britain maybe you have a co pay but ultimately if you're low income or going through a difficult time in terms of your job, the state will help you to pay for dentistry. What that means is that there's a clear state interest in minimizing the cost of people's visits to the dentist. Because fluoridation reduces the rate of cavities which are going to be the most expensive thing to have people get taken care of at the dentist, we tell you that ultimately there's a compelling state interest to put fluoride in the water. A couple of cents up front can save thousands of dollars later on root canals and other dental surgeries. We think that this compelling state interest is enough of a reason to paternalize. Especially because money for health is fungible. Any money that's spent on giving, you know, somebody who has a cavity a new set of teeth, could have been spent on helping a child with some sort of congenital illness. Ultimately we think it's important that we use our money as effectively as possible, that the state is frugal, and fluoridation is certainly that. And the second reason we think you can paternalize is because of the third party harms of not doing so. It may be true that adults can make a choice about whether or not to put fluoride in their water, but children really can't. They can only drink the water that they're given. At that point we think that children who can't choose to consent into this would be doing a lot of damage to their teeth and not rectifying it by using fluoride and ultimately they would suffer in the long term. We think the state needs to intervene to protect them. The third reason we think that we should put fluoride in the water is that it's not an undue burden on anyone. Will tries to tell you that it's unrealistic to ask people who don't want fluoride to drink bottled water. But I think it's an undue burden to ask everybody who wants healthy teeth to go out and buy fluoride so that a couple of hippies don't have to have fluoride in their water. This cuts both ways. We think that at the end of the day, bottled water, in the US at least, is so cheap it's almost free if you buy it in bulk. At that point we don't think it's an undue burden that the tiny minority of people who don't want fluoride have to spend a few dollars every week on water. So at the end of the day we think it's clear that the state should continue to fluoridate water. Thank you.

Potential claims:

1. Fluoridation is effective
2. Fluoridation is a great health achievement
3. Water fluoridation is critical for children
4. Fluoridation is safe
5. Water fluoridation is safe and important to dental health
6. Fluoridation of water is extremely beneficial for citizens, especially children
7. Fluoridation was a worthy project to improve the health
8. Water fluoridation is a safe and effective public health measure
9. Fluoridated water is safe
10. Water fluoridation is effective
11. Water fluoridation is safe, effective

Figure 1: A full example of the annotation task. Given a controversial topic, an argumentative speech discussing it, and a list of potential claims (relevant to the topic and of the same stance as the speech), the goal is determining which claims are mentioned in the speech. To appreciate the difficulty of the task, readers are encouraged to try to annotate this example themselves. The task is described in more detail in §2.

into sentences, it is possible to present a single sentence and a single claim, and ask whether the claim is made or mentioned in the sentence. However, this *sentence-level* setup has three major problems. First, there is a large number of sentence-claim pairs, which makes comprehensive labeling of all pairs unfeasible, even with crowdsourcing. For example, among the 200 speeches of Mirkin et al. (2018) a typical speech contains about 30 sentences, and is labeled vs. 4 claims. Thus, labeling the entire dataset requires labeling some 24,000 pairs. Second, the goal of the annotation process is to provide a fairly comprehensive sample of claims mentioned in speeches (e.g. for training a classifier), yet such pairs are rare. Thus, collecting a sizable amount of such pairs requires labeling a large amount of data. Third, labeling single sentences obscures their context, which may, in some cases, change how they are understood by annotators, thus affecting the collected labels. For example, a claim may not be explicit in a single sentence, but rather implied by a section of the speech.

An alternative to this approach is *speech-level* labeling – presenting an entire speech along with the full list of potential claims. This makes comprehensive labeling of entire speeches feasible, at the cost of added time and complexity. Annotation of a single speech takes at least several minutes of reading and/or listening, and long lists of claims often require iterating over the speech multiple times, since it is hard to memorize its full content in a single pass. It is tempting for an annotator who is not skilled at such tasks to only glimpse through the long text, rather than read it carefully. Conversely, a small, skilled workforce may be able to deal with a task of this complexity, but large-scale data collection by such a workforce is impractical.

To overcome these challenges, we suggest combining the advantages of both setups. Namely, comprehensive labeling of entire speeches using crowdsourcing. The main issue is to identify and motivate a reliable, skilled crowd workforce which is of sufficient size to perform it on a large scale. Similar works attempted to identify reliable crowd annotators based on their previous work (Ho et al., 2013), or other user characteristics like age or education (Li et al., 2014). Behavioral patterns during the task like scrolling and context switching have also been used to predict user reliability in crowdsourcing platforms (Goyal et al., 2018). Here, we rely on their suitability to our specific task, which requires unique skills like reading and listening comprehension and attention to nuance. During the annotation process, we monitor several features of each annotator (see §4), such as agreement with peers and labeling time, and use them to evaluate our confidence in their work. Based on these confidence measures, annotators determined as unreliable are filtered out, and strong ones are retained and rewarded. This monitoring also allowed to identify problems in our task design, which helped in adjusting it to the crowd.

Lastly, annotations from the two annotation schemes are compared, using pairs of claim and speech that were labeled in both (see §5).

The main contributions of this paper are: (i) Presenting a case study of long texts annotation in a complex NLU task, using crowdsourcing; (ii) A detailed description of a mechanism to select annotators that are reliable and qualified to the task using quality control measures taken from their work on our specific data; (iii) An analysis comparing an annotation setup which provides full textual context, to a simpler setup which obscures context information from annotators.

2 The annotation task

Listening comprehension over argumentative content is a new NLU task we recently introduced in Mirkin et al. (2018). This work included a corresponding dataset, annotated by experienced experts. Following is a description of that annotation task, which we now aim to adapt to the crowd.

Each annotation unit is presented in the context of a given controversial topic, such as *we should end water fluoridation*. It is comprised of two parts (see Figure 1): The first is a several-minute long speech, in which a single speaker is arguing for or against the given topic. The speeches are provided in both audio and text, allowing annotators a choice between listening, reading or both. The second part is a list of claims, potentially relevant to the topic and of the same stance as that of the speaker. The objective is identifying the subset of claims mentioned in a given speech. The resulting annotation is a set of speech–claim pairs, in which a pair is considered a *positive* match if the claim is mentioned in the speech (otherwise the pair is considered a *negative* match).

Specifically, annotators were instructed to consider a claim as mentioned in a speech if the statement *"The speaker argued that <claim>"* is true. This statement can be valid even if the speaker was stating the claim using a different phrasing or even if she did not explicitly express the claim, but merely implied it (see Example 1).

The full annotation guidelines are given in the Supplementary Materials.

Example 1 (Claim implied from a speech)
Claim: *Needle exchange reduces the spread of diseases*
Speech: *[...] Without the needle exchange program people are still going to do heroin or other kinds of drugs anyway with dirty or less safe needles. This does lead to things like HIV getting transmitted, it leads to other diseases as well, being more likely to get transmitted [...]*

3 Sentence-level annotation

In a *sentence-level* annotation scheme, the speech text is first split into sentences[2]. Then, pairs of sentence and claim are presented to annotators, who answer whether the claim is stated in the sentence. Figure 2 shows a screenshot of one annotation unit in this scheme. The questions are short, which is advantageous for crowdsourcing, and the collected answers indicate, in addition to whether a claim was mentioned in a speech, *where* was it mentioned, which is potentially important information for methods aimed at automatically identifying claims in speeches.

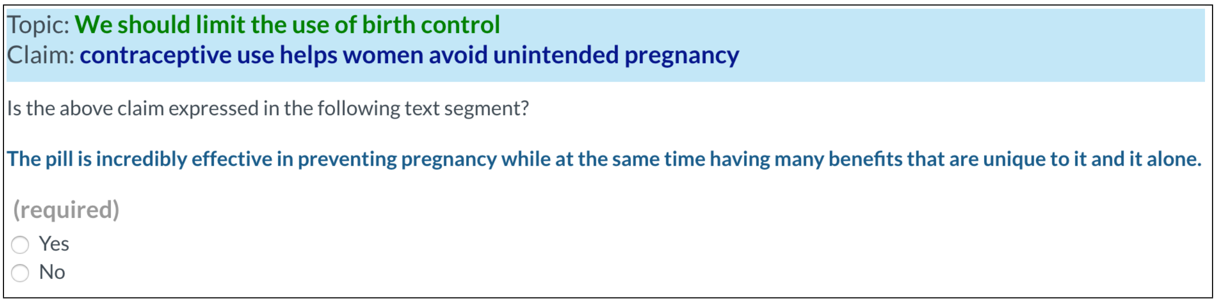
However, this scheme has three major limitations:

– *Scalability*: Comprehensive labeling of all possible sentence-claim pairs is not feasible, even for crowdsourcing. A speech in our data contains, on average, 28.7 sentences, and has 65.6 claims which require annotation. This means having 1,882 claim and sentence pairs for each speech, and sums up to more than 2 million pairs for our data of 1,127 speeches.

A naive approach for reducing the number of pairs which require annotation is randomly sampling sentences from a given speech. However, because claims mentioned in speeches are typically mentioned only once or twice, such sampling would likely miss the mentioning sentences.

Another option is detecting sentences which are semantically similar to the claim, and annotating those with a high similarity. We tried doing so by using *word2vec* (Mikolov et al., 2013): a vector representation for a claim or a sentence was defined as the weighted-average of the vec-

[2] Using a manually created transcription of the audio into text, which includes sentence segmentation.


Topic: We should limit the use of birth control
Claim: contraceptive use helps women avoid unintended pregnancy

Is the above claim expressed in the following text segment?

The pill is incredibly effective in preventing pregnancy while at the same time having many benefits that are unique to it and it alone.

(required)

○ Yes
○ No

Figure 2: A screenshot of one unit within a *sentence-level* annotation scheme, including one claim-sentence pair.

tor representation of its words (using idf weights based on Wikipedia). The similarity between a claim and a sentence was then calculated using the cosine similarity between their vector representations. This increased the fraction of positive pairs, yet introduced a bias: pairs with definite lexical overlap were selected for labeling, but pairs where the claim is paraphrased or implicit were overlooked. Other selection options are possible, but they would likely introduce bias to the labeling process for similar reasons.

– *Limited context*: Deciding whether the claim is mentioned based on a single sentence can be difficult for two reasons. First, it is often hard to fully understand the speaker's intent when reading a single sentence. The sentence may refer to previous parts of the speech or contain an incomplete train of thought. Second, in many cases, a speaker clearly conveys a claim, yet it is not explicitly mentioned in any single sentence. Example 2 shows a claim expressed across several non-consecutive sentences.

Example 2 (Multi-sentence mentioned claim)

Claim: *Compulsory voting is undemocratic*

Speech: *Democracy is about protecting our rights [...] People have a right to not vote [...] We should respect literally any reason a person might not want to vote [...] We should ensure that that person is not penalized for not voting.*

– *Noisy negatives*: A claim mentioned in one of the speech sentences implies that it is mentioned in the speech, yet the opposite is not necessarily true. A prerequisite to establishing that a claim is *not* mentioned in a speech is its annotation as not mentioned *for every speech sentence*. Even then, it is possible that the claim arises from a combination of multiple sentences, and that when reviewing the entire speech, it would nonetheless be considered as mentioned. Thus, negative matches obtained in this scheme are a noisy approximation of the actual speech–claim negative examples.

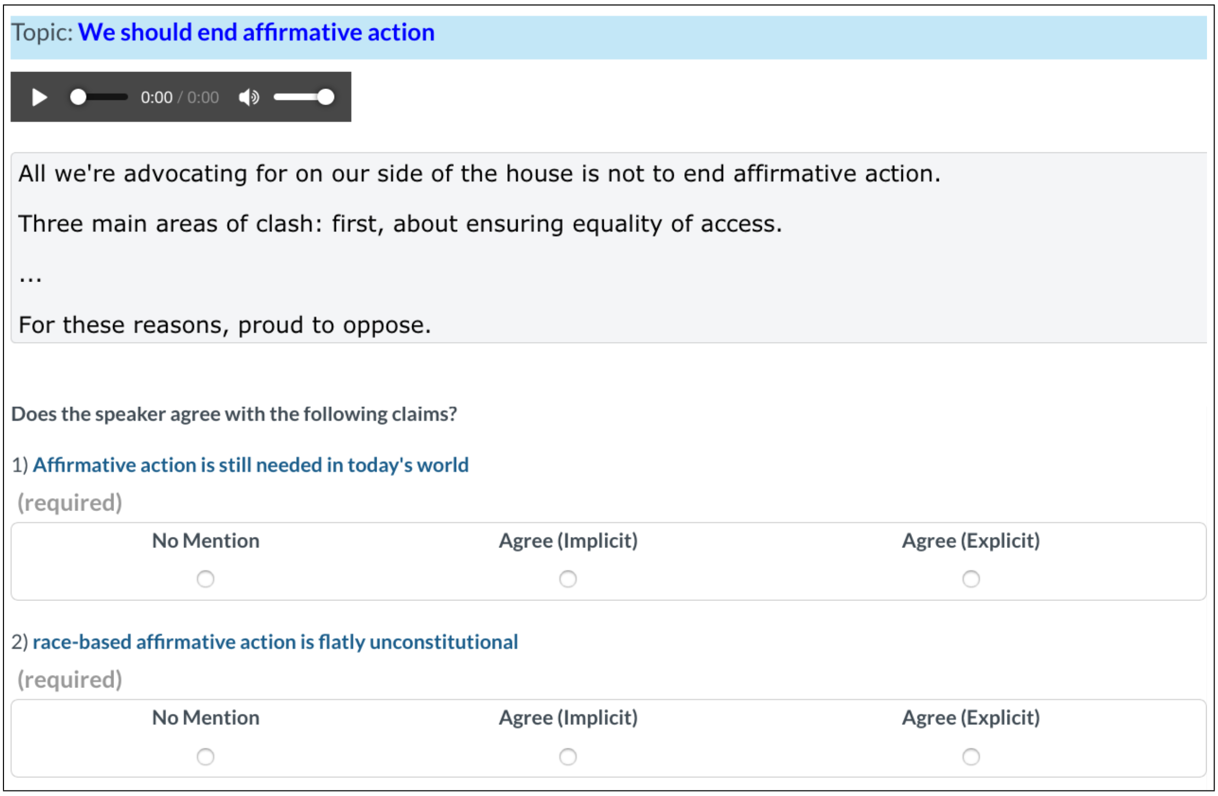
4 Speech-level annotation

The above mentioned limitations of the sentence-level approach suggest that a different setup is desirable. We therefore considered a *speech-level* annotation scheme: annotators were provided with the full speech (text and audio) and a list of at most 20 claims from which they marked those mentioned (Speeches with more than 20 claims were shown more than once). Figure 3 illustrates one annotation unit in this scheme.

The main advantage of this approach is that the full context is available to annotators, making it easier to decide whether a certain idea was expressed. In addition, the collected negative matches are more reliable since annotators access the entire speech. However, this setup does not solve the scalability issue. Each unit is considerably more complex, since it requires the careful evaluation of a long text, while paying attention to nuances and subtleties. Thus, annotating a large volume of data in this scheme is even more challenging, since the common approach for scaling an annotation, namely the use of crowd, is typically applied to short, simple tasks.

Next, we experiment with this scheme using 3 different groups of annotators, using four measures: average pairwise kappa, fraction of high-agreement pairs, fraction of low-agreement pairs and fraction of positive pairs.

Average pairwise kappa is defined by first identifying annotators having at least 5 peers from their group with more than 20 common answers, and averaging their Cohen's Kappa score (Cohen, 1960) with each peer meeting these criteria. Then, the average over annotators is taken as the measure for the group. We note that the applicability of agreement measures like Cohen's Kappa to the crowd has been questioned, in particular for tasks



Topic: We should end affirmative action

All we're advocating for on our side of the house is not to end affirmative action.

Three main areas of clash: first, about ensuring equality of access.

...

For these reasons, proud to oppose.

Does the speaker agree with the following claims?

1) Affirmative action is still needed in today's world

(required)

No Mention	Agree (Implicit)	Agree (Explicit)
○	○	○

2) race-based affirmative action is flatly unconstitutional

(required)

No Mention	Agree (Implicit)	Agree (Explicit)
○	○	○

Figure 3: A screenshot of one unit within a *speech-level* annotation scheme. The unit contains a full speech (the full text is not shown due to space constraints) and a list of claims (partially shown).

within the argumentation domain (Passonneau and Carpenter, 2014; Habernal and Gurevych, 2016). Yet, while their exact value may be of limited interest, using them comparatively allows us to assess the reliability of results from different settings.

High-agreement and *Low-agreement* speech–claim pairs are defined by first defining the label of a pair as the majority vote of the annotators. If this majority includes at least 80% the of annotators, the pair is a *High agreement* pair. If it includes at most 60% of annotators, it is a *low agreement* pair.

The last measure, the *fraction of positive-labeled pairs*, is expected to be similar for different groups of annotators. Additionally, it provides information about the usefulness of the collected data, since a sizable fraction of positive examples is required to allow the development of algorithms which automatically detect claims mentioned in speeches.

4.1 Experts

The first group included highly proficient English-speakers with previous experience in various NLP annotation projects done by our team. Each speech was annotated by five experts.

This step was performed for two reasons: First, to verify that achieving high confidence annotation of our data is feasible, by comparing the annotation measures computed here to those reported in previous similar work which utilized experts. Second, establishing these measures for the experts group creates a baseline for comparison to the measures of crowd-based groups.

Results The *Experts* column of Table 1 summarizes the annotation statistics and results. The inter-annotator agreement of the experts group is 0.4, which is comparable yet somewhat lower, than the value of 0.52 reported in Mirkin et al. (2018). This could be attributed to the different nature of our claims, and having a more skewed data distribution: 20% of our claims are annotated as mentioned, while in the annotation of Mirkin et al. (2018) almost 40% of the claims are so.

4.2 General crowd

As mentioned above, despite having annotated a fairly large number of speech–claim pairs using experts, their limited pace, and the large volume of data, make it impractical to annotate the speeches en-masse in this way. We therefore resorted to the

	Experts	Crowd	Channel
Num. speeches	397	939	1127
Avg. claims per speech	22.8	27.3	65.6
Num. annotated pairs	9,052	25,634	73,931
Num. annotators	14	211	28
Avg. pairwise kappa	0.4	0.24	0.45
High-agreement pairs	80%	67%	68%
Low-agreement pairs	20%	15%	15%
Positive pairs	20%	17%	25%

Table 1: Speech-level annotation statistics (top) and results (bottom), comparing the use of 3 different groups of annotators. The crowd custom channel allowed the annotation of more than 7 times the amount of data annotated by experts, while maintaining quality.

use of the *Figure-Eight*[3] (F8) crowdsourcing platform.

This platform has several built-in quality control mechanisms. Each annotator has a *level*, based on her previous work on the platform. In addition, it encourages the use of *Test Questions* (TQs), questions whose answers are defined by the task's designer, and which are included in a preliminary quiz and in random locations throughout the task. The accuracy of each annotator is then measured on the TQs, and only those who maintain a high accuracy are assigned further questions (those who do not are denied access and their past work is discarded). While the annotators do not know which questions are TQs beforehand, once they submit their answers to one, the F8 platform reveals its correct answers. This allows annotators to review and learn from their mistakes, but also to recognize TQs after their answer was processed.

To create TQs for our task, speech–claim pairs that were unanimously labeled by the experts were taken, and their selected answer was defined as the correct answer. Recall that a question in our task is composed of a speech and a list of claims, and that one needs to answer, for each claim, whether it was mentioned in the speech. For TQs, we've set a known answer for only some of the claims on the list, and ignored answers to the rest. The annotators' minimal required accuracy was set to 0.75, and those with the lowest F8 *level* were denied access. Payment was set to \$0.5 per speech, and each question required seven annotators.

Results Column *Crowd* in Table 1 shows the agreement and quality measurements of this experiment. The obtained agreement is low compared to expert annotators. Such a significant difference is surprising given the TQ mechanism, which was expected to keep only annotators whose answers are consistent with those of the experts.

Analysis Analyzing the obtained annotations raised two major issues:
– ***Implicit claims***: Focusing on high-agreement claim–speech pairs, 91% of the ones annotated by the crowd were labeled as negative, while the experts only annotated 37% of of their high-agreement pairs as such. A deeper look suggested that a major cause were claims alluded to, but not explicitly stated, in the speech (see Example 1). It seemed that while the experts generally agreed on these cases, the guidelines for the untrained crowd annotators did not fully convey the goal of this task. Thus, we changed the annotation labels for the task from binary to *Explicit*, *Implicit*, *No mention*, and added detailed examples of implicit mentions to the guidelines.
– ***User reliability***: Further validation of a random sample of the data revealed many pairs for which, despite a high agreement, the label was wrong, thus raising concerns regarding the reliability of individual annotators. A possible explanation is that the TQs were identified by some annotators, who then made an effort to properly answer only them. This can happen, for example, when an annotator encounters the same TQ twice, or when annotators share answers to TQs with each other, if they are working as part of a group. While a possible solution is increasing the number of TQs to avoid such repetitions, it is still plausible, especially for returning annotators who work on multiple batches of the same task, to see the same TQ multiple times. Furthermore, it has been shown that in any quality assurance mechanism that is based on a fixed set of gold questions, the inherent size limit of the gold set can be exploited by a group of colluding workers, who can build an inferential system to detect which parts of the job are more likely to be gold questions (Checco et al., 2018).

4.3 Custom crowd

F8 allows manually defining a per-task list of annotators who are allowed access to a task, called a *custom channel*. To address the reliability issues raised in our analysis, annotators for such a channel were selected, based on the following per-

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

annotator measures:
– ***Kappa***: Average pairwise kappa vs. others as described above.
– ***TQ failure***: Percentage of incorrectly labeled speech–claim pairs in TQs. This is a more refined assessment of the performance of individual annotators than the one provided by the platform, because the latter considers a TQ as wrong when it has at least one wrongly marked claim, and we assessed speech-claim pairs in TQs individually.
– ***Accept rate***: Percentage of positively annotated speech-claim pairs. Extreme values may suggest that an annotator is not reading carefully, and is rather choosing the same answer again and again.
– ***Judgment time***: Average annotation time of a speech. This is an estimate provided by the platform, and it helps to identify extreme outliers, which do not carefully review the task.
– ***Max pairwise kappa***: The maximal pairwise kappa measured between an annotator and one of her peers. A very high agreement between two annotators suggests that their answers may be coordinated. It may even be a single person, using different ids to access the same task multiple times.
– ***Shared IP***: Whether the annotator's IP address is shared with others doing the same task. Having the same IP address does not imply a single end-user, but it rasies the possibility that it is, or that the end-user is part of a group which may share answers to TQs.

Using these measures, each annotator is assigned a *Reliability Level*:
– **Unreliable**: Annotators who meet at least one of the following conditions: (i) *Accept rate* $< 5\%$ or $> 95\%$; (ii) *Max pairwise kappa* > 0.9; (iii) *Judgment time* < 1 minute; (iv) *shared IP* is true.
– **Low-Quality**: *Kappa* < 0.1 or *TQ failure* $> 50\%$. These are annotators with low quality of work but they are not necessarily malicious users.
– **Reliable**: the rest of the annotators.

The thresholds for the different reliability levels were manually defined after reviewing and analysing the annotation of workers comparing to their obtained scores.

To assess the reliability of the general crowd, these measures were calculated from their annotations, and a *Reliability Level* was assigned to each annotator. Of the 211 annotators who took part in that stage, only 86 were categorized as ***Reliable***. Of all 125 ***Unreliable*** annotators, 50 were also considered ***Low-Quality***. It is possible that the high rate of ***Unreliable*** annotators was due to the complexity of the task which discouraged serious and thorough work, combined with the high payment which attracted many annotators to try it.

We therefore hand-picked a group of ***Reliable*** annotators who contributed the largest number of high quality annotations to be included in a custom channel. By continuing to release in parallel more tasks to the general crowd, this channel was iteratively expanded, knowing such tasks will attract some ***Unreliable*** users, but also more ***Reliable*** ones. Once a task was complete, we calculated annotator levels, and picked new users from those identified as ***Reliable***. Answers from other annotators were discarded. At the same time, we released tasks limited to the custom channel, monitoring annotator performance using the same method.

Notably, when working with the custom channel we disabled the built-in TQ mechanism for two reasons. First, since channel annotators already proved reliable, the quiz given before each batch of the task was no longer necessary. Second, working with TQs technically requires including at least two speeches in every page of the task shown to the annotators (one speech being the TQ). Annotators pointed out that having this configuration makes it harder to focus.

To keep a measure of quality, one or two claims with a known clear answer were embedded as questions for each speech. For example, such a claim might be of a stance opposing that of the speaker, and is thus unlikely to be claimed. We refer to this quality measure as *Hidden Test Questions (HTQ)*, since in contrast to TQs, annotators can't identify them, and they don't know when they erred on them. Annotators only knew their work was closely monitored; and for our internal monitoring an ***HTQ failure*** measure replaces ***TQ failure*** when assessing the custom channel's work.

Results After several iterations, we assembled a group of 28 annotators which achieved similar agreement to that of the expert annotators (see column *Channel* in Table 1), working at a much higher pace. This was probably due to the group including twice as many members as the expert annotators, as well as not being burdened with other annotation tasks (at least not by our team). To keep them motivated, we regularly paid bonuses to annotators based on the quantity and quality of their annotations. The annotators also provided occasional feedback on their experience which helped further improve the design of our task.

To demonstrate the resulting annotation, and to facilitate a basis for algorithms addressing this claim-detection task, an annotation of the speeches from Mirkin et al. (2018) will be made available on our website[4].

5 Comparing the annotations

Having constructed the speech-level annotated dataset, we now revisit our assumption that the simpler sentence-level annotation cannot capture the full context required to correctly label claims in speeches. We compare the annotation of 1,003 claims in 379 speeches via our speech-level methodology with that of the same claims via our initial sentence-level scheme. The latter was done on selected sentences from each speech - those semantically similar to the given claim (see §3).

Table 2 compares labels from both setups. Sentence-level labels are derived from 5,189 sentence–claim pairs (average of 1.7 sentences per speech-claim pair), considering a speech–claim pair positive if the claim was positive in at least one of the sentences annotated for this speech.

The rate of positive pairs is higher in the speech-level scheme: 1,024 pairs (20%) were labeled as positive (explicit or implicit) while only 389 (7.5%) were positive when deriving the label from the sentence-level scheme. As expected, the majority (74%) of sentence-level positives were also considered speech-level positive. Also, 28% of sentence-level negatives were in fact identified as speech-level *positives*, with a high rate of implicitly mentioned claims. Analyzing a sample of such cases suggested that usually the claim can not be pinpointed to a single sentence, but rather arises from a combination of several sentences, while it is also common for the sentence-level annotation to miss the relevant sentence, when one does exist.

Surprisingly, 102 pairs were labeled as positive in the sentence-level but were negative in the speech-level. This is unexpected because a claim that was mentioned in a single sentence of the speech was obviously mentioned in it. Analysis of these pairs revealed that in the majority of them (78%) the sentence-level label was wrong, that is, the claim was not mentioned in the suggested sentence. In many cases it seems that the mistake was due to misinterpretation of the sentence without its context. This confirms the importance of providing a broader context in our task.

Speech / Sent.	Explicit	Implicit	No mention
Positive	150	137	102
Negative	301	436	1,889

Table 2: A comparison of speech-level labels (Explicit, Implicit, No mention) to sentence-level based labels: a Positive claim is one which is positive for least one of the labeled sentences; a Negative claim is one which is negative for all labeled sentences. Note that given a speech, not all of its sentences are labeled, leading to the label mismatches presented here. For further details, see §5.

6 Conclusions and Future Work

We addressed the annotation of claims in argumentative content through crowdsourcing. Due to its complexity, it is not clear that such annotation can be decomposed into simpler sub-tasks in a way that leads to an effective and comprehensive solution. Indeed, our results demonstrate that approximating the full-text context by simple *word2vec*-based sampling of ostensibly-relevant sentences is not sufficient.

Conversely, we show how careful employment of crowdsourcing can address the full, complex problem. By using a combination of various quality control measures to select highly skilled and motivated annotators, we were able to create a committed reliable workforce. This allowed us to obtain large-scale, high quality annotations despite the inherent complexity and subjectivity of this demanding NLU task. We learned that even with a relatively small group of crowd annotators, it is possible to benefit from the advantages of the crowd, namely high pace and scale.

We believe the key to the success of this annotation project was the ongoing learning and improvement we made during the process: analyzing common mistakes directed us to the easier 3-label setup, as well as improve the guidelines to clarify repeating issues and interesting edge cases; keeping an open dialog with our custom channel allowed us to learn from their feedback, and make changes that improved their experience like discarding the TQ mechanism; rewarding good annotators with extra payments made them feel their work is valued and kept them committed to our task.

[4] https://www.research.ibm.com/haifa/dept/vst/debating_data.shtml

In the context of more common NLU tasks, such as those in Wang et al. (2018), our task seems to require an exceptionally high level of language understanding by an automated system seeking to perform it. Since the claims may be implicit in the text, combining the understanding of numerous sentences may be required to perform it adequately. Moreover, if a claim is relevant to the motion, but nonetheless not mentioned in the speech, it may be quite challenging for an automatic system to deduce that such a plausible claim is in fact not implied anywhere in the speech. Hence this task is in line with the motivation of Wang et al. (2019) - a task where there is likely much headroom for an automated system to improve before it reaches human capabilities.

In future work, this dataset could be used to build classifiers of a more global nature, where each labeled speech–claim pair is considered a single unit of information.

Furthermore, speech-level annotation can help facilitate an efficient collection of claim–sentence labels, by first choosing claims labeled as positive in speeches, and annotating them against all speech sentences. Such labels may prove useful in the development of classifiers for identifying claims in single sentences. This method may be useful for other NLU tasks which involve long texts, e.g. Question Answering from long texts.

7 Acknowledgments

We thank George Taylor and the entire Figure-Eight team for their valuable advice and continuous support, which made this annotation project successful. We are thankful to all the debaters and annotators who took part in the creation of this dataset.

References

Alessandro Checco, Jo Bates, and Gianluca Demartini. 2018. All that glitters is goldan attack scheme on gold questions in crowdsourcing. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

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

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tanya Goyal, Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2018. Your behavior signals your reliability: Modeling crowd behavioral traces to ensure quality relevance annotations. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Ivan Habernal and Iryna Gurevych. 2016. 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)*, volume 1, pages 1589–1599.

Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. 2013. Adaptive task assignment for crowdsourced classification. In *International Conference on Machine Learning*, pages 534–542.

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

Hongwei Li, Bo Zhao, and Ariel Fuxman. 2014. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 165–176, New York, NY, USA. ACM.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards cross-lingual textual entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 321–324. Association for Computational Linguistics.

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

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

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 670–679. Association for Computational Linguistics.

Matan Orbach, Yonatan Bilu, Ariel Gera, Yoav Kantor, Lena Dankin, Tamar Lavee, Lili Kotlerman, Shachar Mirkin, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. A dataset of general-purpose rebuttal. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

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

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866.

Merel Scholman and Vera Demberg. 2017. Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 24–33, Valencia, Spain. Association for Computational Linguistics.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. 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*, pages 353–355.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31.

Rui Wang and Chris Callison-Burch. 2010. Cheap facts and counter-facts. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazons Mechanical Turk*, pages 163–167.

Naomi Zeichner, Jonathan Berant, and Ido Dagan. 2012. Crowdsourcing inference-rule evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 156–160. Association for Computational Linguistics.