Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, Dan Roth
Department of Computer and Information Science, University of Pennsylvania
{sihaoc,danielkh,wenpeng,ccb,danroth}@cis.upenn.edu

Abstract

One key consequence of the information revolution is a significant increase and a contamination of our information supply. The practice of fact-checking won't suffice to eliminate the biases in text data we observe, as the degree of factuality alone does not determine whether biases exist in the spectrum of opinions visible to us. To better understand controversial issues, one needs to view them from a diverse yet comprehensive set of *perspectives*.

For example, there are many ways to respond to a *claim* such as *"animals should have lawful rights"*, and these responses form a spectrum of perspectives, each with a *stance* relative to this claim and, ideally, with evidence supporting it. Inherently, this is a natural language understanding task, and we propose to address it as such. Specifically, we propose the task of *substantiated perspective discovery* where, given a *claim*, a system is expected to discover a *diverse* set of *well-corroborated perspectives* that take a *stance* with respect to the claim. Each perspective should be substantiated by *evidence* paragraphs which summarize pertinent results and facts.

We construct PERSPECTRUM, a dataset of claims, perspectives and evidence, making use of online debate websites to create the initial data collection, and augmenting it using search engines in order to expand and diversify our dataset. We use crowdsourcing to filter out noise and ensure high-quality data. Our dataset contains $1k$ claims, accompanied by pools of $10k$ and $8k$ perspective sentences and evidence paragraphs, respectively. We provide a thorough analysis of the dataset to highlight key underlying language understanding challenges, and show that human baselines across multiple subtasks far outperform machine baselines built upon state-of-the-art NLP techniques. This poses a challenge and an opportunity for the NLP community to address.

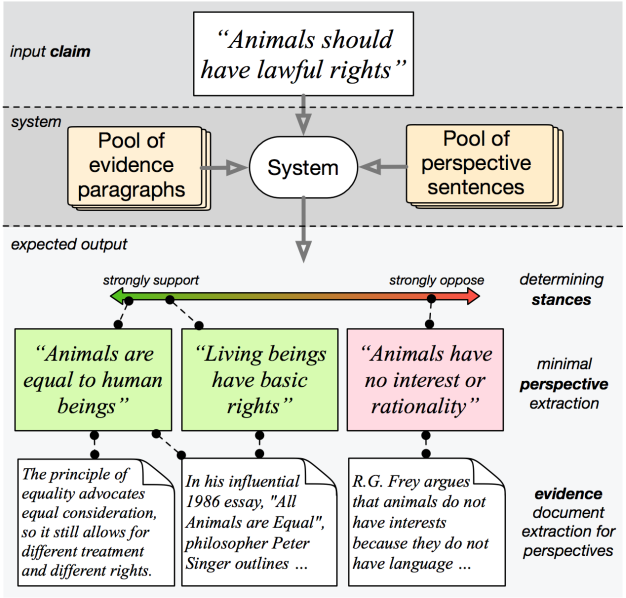

Figure 1: Given a *claim*, a hypothetical system is expected to discover various *perspectives* that are substantiated with *evidence* and their *stance* with respect to the claim.

1 Introduction

Understanding most nontrivial *claim*s requires insights from various *perspective*s. Today, we make use of search engines or recommendation systems to retrieve information relevant to a claim, but this process carries multiple forms of *bias*. In particular, they are optimized relative to the claim (query) presented, and the popularity of the relevant documents returned, rather than with respect to the diversity of the *perspective*s presented in them or whether they are supported by evidence.

In this paper, we explore an approach to mitigating this *selection bias* (Heckman, 1979) when studying (disputed) claims. Consider the *claim* shown in Figure 1: *"animals should have lawful rights."* One might compare the biological similarities/differences between humans and other an-

imals to support/oppose the claim. Alternatively, one can base an argument on morality and rationality of animals, or lack thereof. Each of these arguments, which we refer to as *perspectives* throughout the paper, is an opinion, possibly conditional, in support of a given *claim* or against it. A *perspective* thus constitutes a particular attitude towards a given *claim*.

Natural language understanding is at the heart of developing an ability to identify diverse perspectives for claims. In this work, we propose and study a setting that would facilitate discovering *diverse perspectives* and their supporting evidence with respect to a given *claim*. Our goal is to identify and formulate the key NLP challenges underlying this task, and develop a dataset that would allow a systematic study of these challenges. For example, for the claim in Figure 1, multiple (non-redundant) perspectives should be retrieved from a pool of perspectives; one of them is *"animals have no interest or rationality"*, a *perspective* that should be identified as taking an *opposing* stance with respect to the *claim*. Each *perspective* should also be well-supported by *evidence* found in a pool of potential pieces of evidence. While it might be impractical to provide an exhaustive spectrum of ideas with respect to a *claim*, presenting a small but diverse set of *perspectives* could be an important step towards addressing the *selection bias* problem. Moreover, it would be impractical to develop an exhaustive pool of evidence for all perspectives, from a diverse set of credible sources. We are not attempting to do that. We aim at formulating the core NLP problems, and developing a dataset that will facilitate studying these problems from the NLP angle, realizing that using the outcomes of this research in practice requires addressing issues such as trustworthiness (Pasternack and Roth, 2010, 2013) and possibly others. Inherently, our objective requires understanding the relations between *perspectives* and *claims*, the nuances in the meaning of various *perspectives* in the context of *claims*, and relations between perspectives and evidence. This, we argue, can be done with a diverse enough, but not exhaustive, dataset. And it can be done without attending to the legitimacy and credibility of sources contributing evidence, an important problem but orthogonal to the one studied here.

To facilitate the research towards developing solutions to such challenging issues, we propose

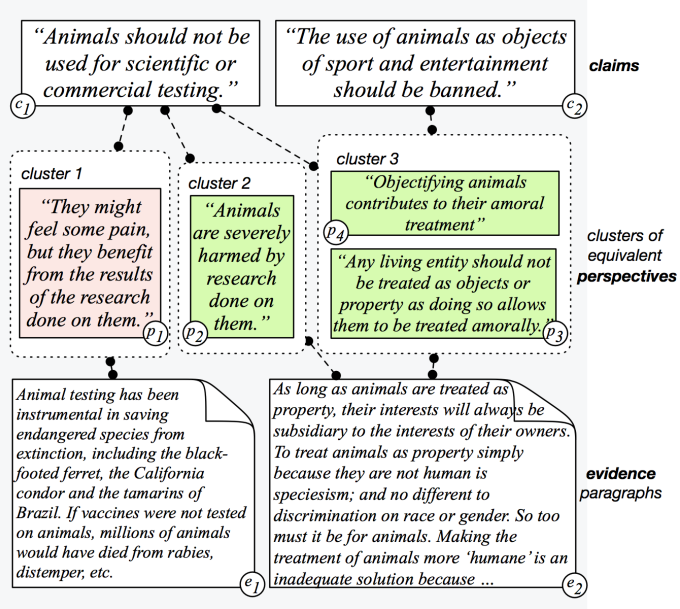

Figure 2: Depiction of a few claims, their *perspectives* and evidences from PERSPECTRUM. The *supporting* and *opposing* perspectives are indicated with green and red colors, respectively.

PERSPECTRUM, a dataset of *claims*, *perspectives* and *evidence* paragraphs. For a given *claim* and pools of *perspectives* and *evidence paragraphs*, a hypothetical system is expected to select the relevant perspectives and their supporting paragraphs.

Our dataset contains 907 claims, 11,164 perspectives and 8,092 evidence paragraphs. In constructing it, we use online debate websites as our initial seed data, and augment it with search data and paraphrases to make it richer and more challenging. We make extensive use of crowdsourcing to increase the quality of the data and clean it from annotation noise.

The contributions of this paper are as follows:

- To facilitate making progress towards the problem of *substantiated perspective discovery*, we create a high-quality dataset for this task.[1]
- We identify and formulate multiple NLP tasks that are at the core of addressing the *substantiated perspective discovery* problem. We show that humans can achieve high scores on these tasks.
- We develop competitive baseline systems for each sub-task, using state-of-the-art techniques.

[1] https://github.com/CogComp/perspectrum

2 Design Principles and Challenges

In this section we provide a closer look into the challenge and propose a collection of tasks that move us closer to *substantiated perspective discovery*. To clarify our description we use to following notation. Let c indicate a target claim of interest (for example, the claims c_1 and c_2 in Figure 2). Each claim c is addressed by a collection of perspectives $\{p\}$ that are grouped into clusters of *equivalent* perspectives. Additionally, each perspective p is supported, relative to c, by at least one evidence paragraph e, denoted $e \vDash p|c$.

Creating systems that would address our challenge in its full glory requires solving the following interdependent tasks:

Determination of argue-worthy claims: not every claim requires an in-depth discussion of perspectives. For a system to be practical, it needs to be equipped with understanding argumentative structures (Palau and Moens, 2009) in order to discern disputed claims from those with straightforward responses. We set aside this problem in this work and assume that all the inputs to the systems are discussion-worthy claims.

Discovery of pertinent perspectives: a system is expected to recognize argumentative sentences (Cabrio and Villata, 2012) that directly address the points raised in the disputed claim. For example, while the perspectives in Figure 2 are topically related to the claims, p_1, p_2 do not directly address the focus of claim c_2 (i.e., *"use of animals" in "entertainment"*).

Perspective equivalence: a system is expected to extract a *minimal* and *diverse* set of perspectives. This requires the ability to discover equivalent perspectives p, p', with respect to a claim c: $p|c \approx p'|c$. For instance, p_3 and p_4 are equivalent in the context of c_2; however, they might not be equivalent with respect to any other claim. The conditional nature of perspective equivalence differentiates it from the *paraphrasing* task (Bannard and Callison-Burch, 2005).

Stance classification of perspectives: a system is supposed to assess the stances of the perspectives with respect to the given claim (supporting, opposing, etc.) (Hasan and Ng, 2014).

Substantiating the perspectives: a system is expected to find valid evidence paragraph(s) in support of each perspective. Conceptually, this is similar to the well-studied problem of textual entailment (Dagan et al., 2013) except that here the entailment decisions depend on the choice of claims.

3 Related Work

Claim verification. The task of *fact verification* or *fact-checking* focuses on the assessment of the truthfulness of a claim, given evidence (Vlachos and Riedel, 2014; Mitra and Gilbert, 2015; Samadi et al., 2016; Wang, 2017; Nakov et al., 2018; Hanselowski et al., 2018; Karimi et al., 2018; Alhindi et al., 2018). These tasks are highly related to the task of textual-entailment that has been extensively studied in the field (Bentivogli et al., 2008; Dagan et al., 2013; Khot et al., 2018). Some recent work study jointly the problem of identifying evidence and verifying that it supports the claim (Yin and Roth, 2018).

Our problem structure encompasses the *fact verification* problem (as verification of *perspectives* from *evidence*; Figure 1).

Stance classification. Stance classification aims at detecting phrases that *support* or *oppose* a given claim. The problem has gained significant attention in the recent years; to note a few important ones, Hasan and Ng (2014) create a dataset of dataset text snippets, annotated with "reasons" (similar to *perspectives* in this work) and stances (whether they support or oppose the claim). Unlike this work, our pool of the relevant "reasons" is not restricted. Ferreira and Vlachos (2016) create a dataset of rumors (claims) coupled with news headlines and their stances. There are a few other works that fall in this category (Boltužić and Šnajder, 2014; Park and Cardie, 2014; Rinott et al., 2015; Swanson et al., 2015; Mohammad et al., 2016; Sobhani et al., 2017; Bar-Haim et al., 2017). Our approach here is closely related to existing work in this direction, as stance classification is part of the problem studied here.

Argumentation. There is a rich literature on *formalizing* argumentative structures from free text. There are a few theoretical works that lay the ground work to characterizing units of arguments and argument-inducing inference (Teufel et al., 1999; Toulmin, 2003; Freeman, 2011).

Others have studied the problem of extracting argumentative structures from free-form text; for example, Palau and Moens (2009); Khatib et al. (2016); Ajjour et al. (2017) studied elements of arguments and the internal relations between them.

Dataset	Stance Classification	Evidence Verification	Human Verified	Open Domain
PERSPECTRUM (this work)	✓	✓	✓	✓
FEVER (Thorne et al., 2018)	✗	✓	✓	✓
(Wachsmuth et al., 2017)	✓	✓	✗	✓
LIAR (Wang, 2017)	✗	✓	✓	✓
(Vlachos and Riedel, 2014)	✗	✓	✓	✓
(Hasan and Ng, 2014)	✓	✗	✓	✗

Table 1: Comparison of PERSPECTRUM to a few notable datasets in the field.

Feng and Hirst (2011) classified an input into one of the argument schemes. Habernal and Gurevych (2017) provided a large corpus annotated with argument units. Cabrio and Villata (2018) provide a thorough survey the recent work in this direction. A few other works studied other aspects of argumentative structures (Cabrio and Villata, 2012; Khatib et al., 2016; Lippi and Torroni, 2016; Zhang et al., 2017; Stab and Gurevych, 2017).

A few recent works use a similar conceptual design that involves a *claim*, *perspectives* and *evidence*.These works are either too small due to the high cost of construction (Aharoni et al., 2014) or too noisy because of the way they are crawled from online resources (Wachsmuth et al., 2017; Hua and Wang, 2017). Our work makes use of both online content and of crowdsourcing, in order to construct a sizable and high-quality dataset.

4 The PERSPECTRUM Dataset

4.1 Dataset construction

In this section we describe a multi-step process, constructed with detailed analysis, substantial refinements and multiple pilots studies.

We use crowdsourcing to annotate different aspects of the dataset. We used Amazon Mechanical Turk (AMT) for our annotations, restricting the task to workers in five English-speaking countries (USA, UK, Canada, New Zealand, and Australia), more than 1000 finished HITs and at least a 95% acceptance rate. To ensure the diversity of responses, we do not require additional qualifications or demographic information from our annotators.

For any of the annotations steps described below, the users are guided to an external platform where they first read the instructions and try a verification step to make sure they have understood the instructions. Only after successful completion are they allowed to start the annotation tasks.

Throughout our annotations, it is our aim to make sure that the workers are responding objectively to the tasks (as opposed to using their personal opinions or preferences). The screen-shots of the annotation interfaces for each step are included in the Appendix (Section A.3).

In the steps outlined below, we filter out a subset of the data with low rater–rater agreement ρ (see Appendix A.2). In certain steps, we use an information retrieval (IR) system[2] to generate the best candidates for the task at hand.

Step 1: The initial data collection. We start by crawling the content of a few notable debating websites: `idebate.com`, `debatewise.org`, `procon.org`. This yields $\sim 1k$ claims, $\sim 8k$ perspectives and $\sim 8k$ evidence paragraphs (for complete statistics, see Table 4 in the Appendix). This data is significantly noisy and lacks the structure we would like. In the following steps we explain how we denoise it and augment it with additional data.

Step 2a: Perspective verification. For each perspective we verify that it is a complete English sentence, with a clear stance with respect to the given claim. For a fixed pair of *claim* and *perspective*, we ask the crowd-workers to label the perspective with one of the five categories of *support*, *oppose*, *mildly-support*, *mildly-oppose*, or *not a valid perspective*. The reason that we ask for two levels of intensity is to distinguish *mild* or *conditional* arguments from those that express *stronger* positions.

Every 10 claims (and their relevant perspectives) are bundled to form a HIT. Three independent annotators solve a HIT, and each gets paid $1.5-2 per HIT. To get rid of the ambiguous/noisy perspectives we measure rater-rater agreement on the resulting data and retain only the subset which has a significant agreement of $\rho \geq 0.5$. To account for minor disagreements in the intensity of

[2] www.elastic.co

perspective stances, before measuring any notion of agreement, we collapse the five labels into three labels, by collapsing *mildly-support* and *mildly-oppose* into *support* and *oppose*, respectively.

To assess the quality of these annotations, two of the authors independently annotate a random subset of instances in the previous step (328 perspectives for 10 claims). Afterwards, the differences were adjudicated. We measure the accuracy adjudicated results with AMT annotations to estimate the quality of our annotation. This results in an accuracy of 94%, which shows high-agreement with the crowdsourced annotations.

Step 2b: Perspective paraphrases. To enrich the ways the perspectives are phrased, we crowdsource paraphrases of our perspectives. We ask annotators to generate two paraphrases for each of the 15 perspectives in each HIT, for a reward of $1.50.

Subsequently, we perform another round of crowdsourcing to verify the generated paraphrases. We create HITs of 24 candidate paraphrases to be verified, with a reward of $1. Overall, this process gives us $\sim$ 4.5 paraphrased perspectives. The collected paraphrases form clusters of equivalent perspectives, which we refine further in the later steps.

Step 2c: Web perspectives. In order to ensure that our dataset contains more realistic sentences, we use web search to augment our pool of perspectives with additional sentences that are topically related to what we already have. Specifically, we use Bing search to extract sentences that are similar to our current pool of perspectives, by querying "claim+perspective". We create a pool of relevant web sentences and use an IR system (introduced earlier) to retrieve the 10 most similar sentences. These candidate perspectives are annotated using (similar to step 2a) and only those that were agreed upon are retained.

Step 2d: Final perspective trimming. In a final round of annotation for perspectives, an expert annotator went over all the claims in order to verify that all the equivalent perspectives are clustered together. Subsequently, the expert annotator went over the most similar claim-pairs (and their perspectives), in order to annotate the missing perspectives shared between the two claims. To cut the space of claim pairs, the annotation was done on the top 350 most similar claim pairs retrieved by the IR system.

Category	Statistic	Value
Claims	# of claims (step 1)	907
	avg. claim length (tokens)	8.9
	median claims length (tokens)	8
	max claim length (tokens)	30
	min claim length (tokens)	3
Perspectives	# of perspectives	11,164
	Debate websites (step 1)	4,230
	Perspective paraphrase (step 2b)	4,507
	Web (step 2c)	2,427
	# of perspectives with stances	5,095
	# of "support" perspectives	2,627
	# of "opposing" perspectives	2,468
	avg size of perspective clusters	2.3
	avg length of perspectives (tokens)	11.9
Evidences	# of total evidences (step 1)	8,092
	avg length of evidences (tokens)	168

Table 2: A summary of PERSPECTRUM statistics

Step 3: Evidence verification. The goal of this step is to decide whether a given evidence paragraph provides enough substantiations for a perspective or not. Performing these annotations exhaustively for any perspective-evidence pair is not possible. Instead, we make use of a retrieval system to annotate only the relevant pairs. In particular, we create an index of all the perspectives retained from *step 2a*. For a given evidence paragraph, we retrieve the top relevant perspectives. We ask the annotators to note whether a given evidence paragraph *supports* a given perspective or not. Each HIT contains a 20 evidence paragraphs and their top 8 relevant candidate perspectives. Each HIT is paid $1 and annotated by at least 4 independent annotators.

In order to assess the quality of our annotations, a random subset of instances (4 evidence-perspective pairs) are annotated by two independent authors and the differences are adjudicated. We measure the accuracy of our adjudicated labels versus AMT labels, resulting in 87.7%. This indicates the high quality of the crowdsourced data.

4.2 Statistics on the dataset

We now provide a brief summary of PERSPECTRUM. The dataset contains about $1k$ claims with a significant length diversity (Table 2). Additionally, the dataset comes with $\sim$ $12k$ perspectives, most of which were generated through paraphrasing (step 2b). The perspectives which convey the same point with respect to a claim are grouped into clusters. On average, each cluster has a size of 2.3 which shows that, on average, many perspectives have

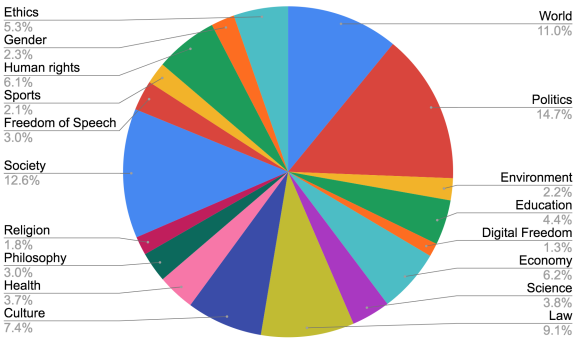

Figure 3: Distribution of claim topics.

equivalents. More granular details are available in Table 2.

To better understand the topical breakdown of claims in the dataset, we crowdsource the set of "topics" associated with each *claim* (e.g., *Law, Ethics, etc.*) We observe that, as expected, the three topics of *Politics, World, and Society* have the biggest portions (Figure 3). Additionally, the included claims touch upon 10+ different topics. Figure 4 depicts a few popular categories and sampled questions from each.

4.3 Required skills

We perform a closer investigation of the abilities required to solve the stance classification task. One of the authors went through a random subset of claim-perspectives pairs and annotated each with the abilities required in determining their stances labels. We follow the common definitions used in prior work (Sugawara et al., 2017; Khashabi et al., 2018). The result of this annotation is depicted in Figure 5. As can be seen, the problem requires understanding of *commonsense*, i.e., an understanding that is commonly shared among humans and rarely gets explicitly mentioned in the text. Additionally, the task requires various types of *coreference* understanding, such as *event coreference* and *entity coreference*.

5 Empirical Analysis

In this section we provide empirical analysis to address the tasks. We create a split of 60%/15%/25% of the data train/dev/test. In order to make sure our baselines are not overfitting to the keywords of each topic (the "topic" annotation from Section 4.2), we make sure to have claims with the same topic fall into the same split.

For simplicity, we define a notation which we will extensively use for the rest of this paper. The clusters of equivalent perspectives are denoted as $[\![p]\!]$, given a representative member p. Let $P(c)$ denote the collection of relevant perspectives to a claim c, which is the union of all the equivalent perspectives participating in the claim: $\{[\![p_i]\!]\}_i$. Let $E([\![p]\!]) = E(p) = \bigcup_i e_i$ denote the set of evidence documents lending support to a perspective p. Additionally, denote the two pools of perspectives and evidence with $\mathcal{U}^p$ and $\mathcal{U}^e$, respectively.

5.1 Systems

We make use of the following systems in our evaluation:

IR (Information Retrieval). This baseline has been successfully used for related tasks like Question Answering (Clark et al., 2016). We create two versions of this baseline: one with the pool of perspectives $\mathcal{U}^p$ and one with the pool of evidences $\mathcal{U}^e$. We use this system to retrieve a ranked list of best matching perspective/evidence from the corresponding index.

BERT (Contextual representations). A recent state-of-the-art contextualized representation (Devlin et al., 2018). This system has been shown to be effective on a broad range of natural language understanding tasks.

Human Performance. Human performance provides us with an estimate of the best achievable results on datasets. We use human annotators to measure human performance for each task. We randomly sample 10 claims from the test set, and instruct two expert annotators to solve each of T1 to T4.

5.2 Evaluation metrics.

We perform evaluations on four different subtasks in our dataset. In all of the following evaluations, the systems are given the two pools of perspectives $\mathcal{U}^p$ and evidences $\mathcal{U}^e$.

T1: Perspective extraction. A system is expected to return the collection of mutually disjoint perspectives with respect to a given claim. Let $\hat{P}(c)$ be the set of output perspectives. Define the precision and recall as $\text{Pre}(c) = \frac{\sum_{\hat{p}\in\hat{P}(c)} \mathbf{1}\{\exists p, s.t. \hat{p}\in[\![p]\!]\}}{|\hat{P}(c)|}$ and $\text{Rec}(c) = \frac{\sum_{\hat{p}\in\hat{P}(c)} \mathbf{1}\{\exists p, s.t. \hat{p}\in[\![p]\!]\}}{|P(c)|}$ respectively. To calculate dataset metrics, the aforementioned per-claim metrics are averaged across all the claims in the test set.

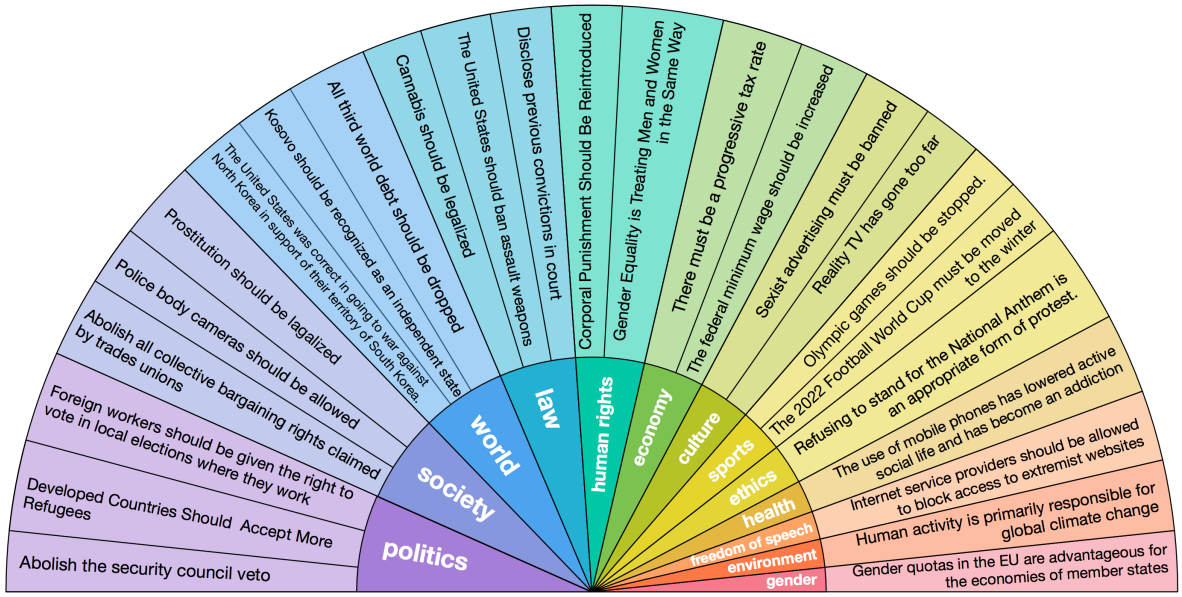

Figure 4: Visualization of the major topics and sample claims in each category.

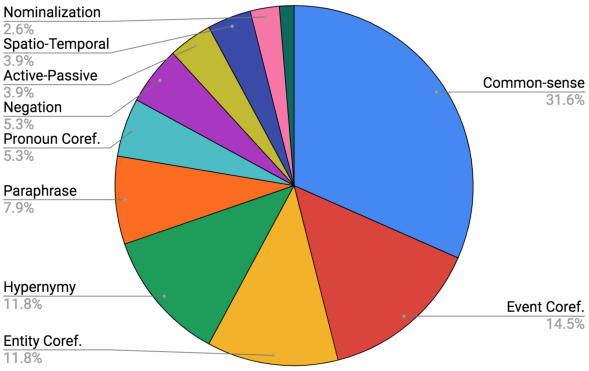

Figure 5: The set of reasoning abilities required to solve the stance classification task.

T2: Perspective stance classification. Given a claim, a system is expected to label every perspective in $P(c)$ with one of two labels *support* or *oppose*. We use the well-established definitions of precision-recall for this binary classification task.

T3: Perspective equivalence. A system is expected to decide whether two given perspectives are equivalent or not, with respect to a given claim. We evaluate this task in a way similar to a clustering problem. For a pair of perspectives $p_1, p_2 \in P(c)$, a system predicts whether the two are in the same cluster or not. The ground-truth is whether there is a cluster which contains both of the perspectives or not: $\exists \tilde{p}\ s.t.\ \tilde{p} \in P(c) \wedge p_1, p_2 \in [\![\tilde{p}]\!]$. We use this pairwise definition for all the pairs in $P(c) \times P(c)$, for any claim c in the test set.

T4: Extraction of supporting evidences. Given a perspective p, we expect a system to return all the evidence $\{e_i\}$ from the pool of evidence $\mathcal{U}^e$. Let $\hat{E}(p)$ and $E(p)$ be the predicted and gold evidence for a perspective p. Define macro-precision and macro-recall as $\text{Pre}(p) = \frac{|\hat{E}(p) \cap E(p)|}{|\hat{E}(p)|}$ and $\text{Rec}(p) = \frac{|\hat{E}(p) \cap E(p)|}{|E(p)|}$, respectively. The metrics are averaged across all the perspectives p participating in the test set.

T5: Overall performance. The goal is to get estimates of the overall performance of the systems. Instead of creating a complex measure that would take all the aspects into account, we approximate the overall performance by multiplying the disjoint measures in $T1$, $T2$ and $T4$. While this gives an estimate on the overall quality, it ignores the pipeline structure of the task (e.g., the propagation of the errors throughout the pipeline). We note that the task of $T3$ (perspective equivalence) is indirectly being measured within $T1$. Furthermore, since we do not report an IR performance on $T2$, we use the "always supp" baseline instead to estimate an overall performance for IR.

5.3 Results

5.3.1 Minimal perspective extraction (T1)

Table 3 shows a summary of the experimental results. To measure the performance of the IR system, we use the index containing $\mathcal{U}^p$. Given each claim, we query the top k perspectives, ranked ac-

cording to their retrieval scores. We tune k on our development set and report the results on the test section according to the tuned parameter. We use IR results as candidates for other solvers (including humans). For this task, IR with top-15 candidates yields >90% recall (for the PR-curve, see Figure 6 in the Appendix). In order to train BERT on this task, we use the IR candidates as the training instances. We then tune a threshold on the dev data to select the top relevant perspectives. In order to measure human performance, we create an interface where two human annotators see IR top-k and select a *minimal* set of perspectives (i.e., no two equivalent perspectives).

5.3.2 Perspective stance classification (T2)

We measure the quality of perspective stance classification, where the input is a claim-perspective pair, mapped to {support, oppose}. The candidate inputs are generated on the collection of perspectives $P(c)$ relevant to a claim c. To have an understanding of a lower bound for the metric, we measure the quality of an always-support baseline. We measure the performance of BERT on this task as well, which is about 20% below human performance. This might be because this task requires a deep understanding of *commonsense* knowledge/reasoning (as indicated earlier in Section 5). Since a retrieval system is unlikely to distinguish perspectives with different stances, we do not report the IR performance for this task.

5.3.3 Perspective equivalence (T3)

We create instances in the form of (p_1, p_2, c) where $p_1, p_2 \in P(c)$. The expected label is whether the two perspectives belong to the same equivalence class or not. In the experiments, we observe that BERT has a significant performance gain of $\sim 36\%$ over the IR baseline. Meanwhile, this system is behind human performance by a margin of $\sim 20\%$.

5.3.4 Extraction of supporting evidence (T4)

We evaluate the systems on the extraction of items from the pool of evidences $\mathcal{U}^e$, given a *claim-perspective* pair. To measure the performance of the IR system working with the index containing $\mathcal{U}^e$ we issue a query containing the concatenation of a perspective-claim pair. Given the sorted results (according to their retrieval confidence score), we select the top candidates using a threshold parameter tuned on the dev set. We also use the IR system's candidates (top-60) for other baselines. This set of candidates yields a >85% recall (for the PR-curve, see Figure 6 in the Appendix). We train BERT system to map each (gold) *claim-perspective* pair to its corresponding *evidence* paragraph(s). Since each evidence paragraph could be long (hence hard to feed into BERT), we split each evidence paragraph into sliding windows of 3 sentences. For each *claim-perspective* pair, we use all 3-sentences windows of gold evidence paragraphs as positive examples, and rest of the IR candidates as negative examples. In the run-time, if a certain percentage (tuned on the dev set) of the sentences from a given evidence paragraph are predicted as positive by BERT, we consider the whole evidence as positive (i.e. it supports a given *perspective*).

Overall, the performances on this task are lower, which could probably be expected, considering the length of the evidence paragraphs. Similar to the previous scenarios, the BERT solver has a significant gain over a trivial baseline, while standing behind human with a significant margin.

Setting	Target set	System	*Pre.*	*Rec.*	*F*1
T1: Perspective relevance	$\mathcal{U}^p$	IR	46.8	34.9	40.0
		IR + BERT	47.3	54.8	**50.8**
		IR + Human	63.8	83.8	72.5
T2: Perspective stance	$P(c)$	Always "supp."	51.6	100.0	68.0
		BERT	70.5	71.1	**70.8**
		Human	91.3	90.6	90.9
T3: Perspective equivalence	$P(c)^2$	Always "¬equiv."	100.0	11.9	21.3
		Always "equiv."	20.3	100.0	33.7
		IR	36.5	36.5	36.5
		BERT	85.3	50.8	**63.7**
		Human	87.5	80.2	83.7
T4: Evidence extraction	$\mathcal{U}^e$	IR	42.2	52.5	46.8
		IR + BERT	69.7	46.3	**55.7**
		IR + Human	70.8	53.1	60.7
T5: Overall	$\mathcal{U}^p$,$\mathcal{U}^e$	IR	-	-	12.8
		IR + BERT	-	-	**17.5**
		IR + Human	-	-	40.0

Table 3: Quality of different baselines on different subtasks (Section 5). All the numbers are in percentage. Top machine baselines are in **bold**.

6 Discussion

As one of the key consequences of the information revolution, *information pollution* and *over-personalization* have already had detrimental effects on our life. In this work, we attempt to facil-

itate the development of systems that aid in better organization and access to information, with the hope that the access to more diverse information can address over-personalization too (Vydiswaran et al., 2014).

The dataset presented here is not intended to be *exhaustive*, nor does it attempt to reflect a true distribution of the important claims and perspectives in the world, or to associate any of the perspective and identified evidence with levels of expertise and trustworthiness. Moreover, it is important to note that when we ask crowd-workers to evaluate the validity of perspectives and evidence, their judgement process can potentially be influenced by their prior beliefs (Markovits and Nantel, 1989). To avoid additional biases introduced in the process of dataset construction, we try to take the least restrictive approach in filtering dataset content beyond the necessary quality assurances. For this reason, we choose not to explicitly ask annotators to filter contents based on the intention of their creators (e.g. offensive content).

A few algorithmic components were not addressed in this work, although they are important to the complete *perspective discovery and presentation* pipeline. For instance, one has to first verify that the input to the system is a reasonably well-phrased and an argue-worthy claim. And, to construct the pool of perspectives, one has to extract relevant arguments (Levy et al., 2014). In a similar vein, since our main focus is the study of the relations between *claim*s, *perspective*s, and *evidence*, we leave out important issues such as their degree of factuality (Vlachos and Riedel, 2014) or trustworthiness (Pasternack and Roth, 2014, 2010) as separate aspects of problem.

We hope that some of these challenges and limitations will be addressed in future work.

7 Conclusion

The importance of this work is three-fold; we define the problem of *substantiated perspective discovery* and characterize language understanding tasks necessary to address this problem. We combine online resources, web data and crowdsourcing and create a high-quality dataset, in order to drive research on this problem. Finally, we build and evaluate strong baseline supervised systems for this problem. Our hope is that this dataset would bring more attention to this important problem and would speed up the progress in this direction.

There are two aspects that we defer to future work. First, the systems designed here assumed that the input are valid claim sentences. To make use of such systems, one needs to develop mechanisms to recognize valid argumentative structures. In addition, we ignore trustworthiness and credibility issues, important research issues that are addressed in other works.

Acknowledgments

The authors would like to thank Jennifer Sheffield, Stephen Mayhew, Shyam Upadhyay, Nitish Gupta and the anonymous reviewers for insightful comments and suggestions. This work was supported in part by a gift from Google and by Contract HR0011-15-2-0025 with the US Defense Advanced Research Projects Agency (DARPA). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. 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*, pages 64–68.

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit Segmentation of Argumentative Texts. In *Proceedings of the Fourth Workshop on Argument Mining (ArgMining 2017)*, pages 118–128.

Tariq Alhindi, Savvas Petridis, Smar Muresan, and a. 2018. Where is your Evidence: Improving Fact-checking by Justification Modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. 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*, volume 1, pages 251–261.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2008. The Sixth PASCAL Recognizing Textual Entailment Challenge. In *TAC*.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

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

Elena Cabrio and Serena Villata. 2018. Five Years of Argument Mining: a Data-driven Analysis. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5427–5433.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzoto. 2013. Recognizing textual entailment: Models and applications.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics*.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 987–996. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

James B Freeman. 2011. *Argument Structure:: Representation and Theory*, volume 18. Springer Science & Business Media.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018. A Retrospective Analysis of the Fake News Challenge Stance Detection Task. *arXiv preprint arXiv:1806.05180*.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the ACL (Volume 1: Long Papers)*.

James J Heckman. 1979. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

Xinyu Hua and Lu Wang. 2017. Understanding and Detecting Supporting Arguments of Diverse Types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 203–208.

Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. 2018. Multi-Source Multi-Class Fake News Detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1546–1557.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Conference of the North American Chapter of the Association for Computational Linguistics (Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL))*.

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *Proceedings of AAAI*.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Marco Lippi and Paolo Torroni. 2016. Argument Mining from Speech: Detecting Claims in Political Debates. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 2979–2985.

Henry Markovits and Guilaine Nantel. 1989. The belief-bias effect in the production and evaluation of logical conclusions. *Memory and Cognition*, 17(1):11–17.

Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations. In *ICWSM*, pages 258–267. AAAI Press.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372–387. Springer.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38.

Jeff Pasternack and Dan Roth. 2010. Knowing what to believe (when you already know something). In *Proc. of the International Conference on Computational Linguistics (COLING)*, Beijing, China.

Jeff Pasternack and Dan Roth. 2013. Latent credibility analysis. In *Proc. of the International World Wide Web Conference (WWW)*.

Jeff Pasternack and Dan Roth. 2014. Judging the veracity of claims and reliability of sources with fact-finders. In Anwitaman Datta Xin Liu and Ee-Peng Lim, editors, *Computational Trust Models and Machine Learning*, pages 39–72. Chapman and Hall/CRC.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. 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*, pages 440–450.

Mehdi Samadi, Partha Pratim Talukdar, Manuela M Veloso, and Manuel Blum. 2016. ClaimEval: Integrated and Flexible Framework for Claim Evaluation Using Credibility of Sources. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 222–228.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 551–557.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. 2017. Prerequisite skills for reading comprehension: Multi-perspective analysis of mctest datasets and systems. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 809–819.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22.

V.G.Vinod Vydiswaran, Chengxiang Zhai, Dan Roth, and Peter Pirolli. 2014. Overcoming bias to learn about controversial topics. *Journal of the American Society for Information Science and Technology (JASIST)*.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59.

William Yang Wang. 2017. " Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 422–426.

Wenpeng Yin and Dan Roth. 2018. Twowingos: A two-wing optimization strategy for evidential claim verification. In *EMNLP*.

Fan Zhang, Homa B Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1568–1578.

A Supplemental Material

A.1 Statistics

We provide brief statistics on the sources of different content in our dataset in Table 4. In particular, this table shows:

1. the size of the data collected from online debate websites (step 1).
2. the size of the data filtered out (step 2a).
3. the size of the perspectives added by paraphrases (step 2b).
4. the size of the perspective candidates added by web (step 2c).

	Website	# of claims	# of perspectives	# of evidences
after step 1	idebate	561	4136	4133
	procon	50	960	953
	debatewise	395	3039	3036
	total	1006	8135	8122
after step 2a	idebate	537	2571	–
	procon	49	619	–
	debatewise	361	1462	–
	total	947	4652	–
step 2b	paraphrases	–	4507	–
step 2c	web perspectives	–	2427	–

Table 4: The dataset statistics (See section 4.1).

A.2 Measure of agreement

We use the following definition formula in calculation of our measure of agreement. For a fixed subject (problem instance), let n_j represent the number of raters who assigned the given subject to the j-th category. The measure of agreement is defined as

$$\rho \triangleq \frac{1}{n(n-1)} \sum_{j=1}^{k} n_j(n_j - 1)$$

where for $n = \sum_{j=1}^{k} n_j$. Intuitively, this function measure concentration of values the vector $(n_1, ..., n_k)$. Take the edge cases:

- Values concentrated: $\exists j, n_j = n$ (in other words $\forall i \neq j, n_i = 0) \Rightarrow P = 1.0$.
- Least concentration (uniformly distribution): $n_1 = n_2 = ... = n_k \Rightarrow \rho = 0.0$.

This definition is used in calculation of more extensive agreement measures (e.g, Fleiss' kappa (Fleiss and Cohen, 1973)). There multiple ways of interpreting this formula:

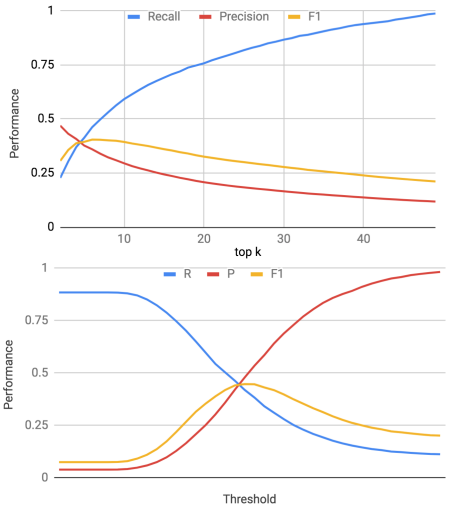

Figure 6: Candidates retrieved from IR baselines vs Precision, Recall, F1, for T1 and T4 respectively.

- It indicates how many rater–rater pairs are in agreement, relative to the number of all possible rater–rater pairs.
- One can interpret this measure by a simple combinatorial notions. Suppose we have sets $A_1, ...A_k$ which are pairwise disjunct and for each j let $n_j = |A_j|$. We choose randomly two elements from $A = A_1 \cup A_2 \cup ... \cup A_k$. Then the probability that they are from the same set is the expressed by ρ.
- We can write ρ in terms of $\sum_{i=1}^{k}(n_i - n/k)^2/(n/k)$ which is the conventional *Chi-Square statistic* for testing if the vector of n_i values comes from the all-categories-equally-likely flat multinomial model.

A.3 crowdsourcing interfaces

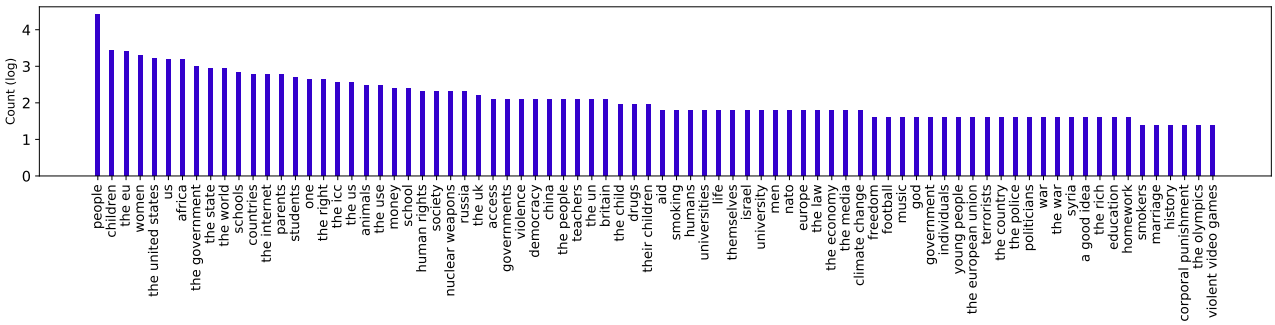

Figure 7: Histogram of popular noun-phrases in our dataset. The y-axis shows count in logarithmic scale.

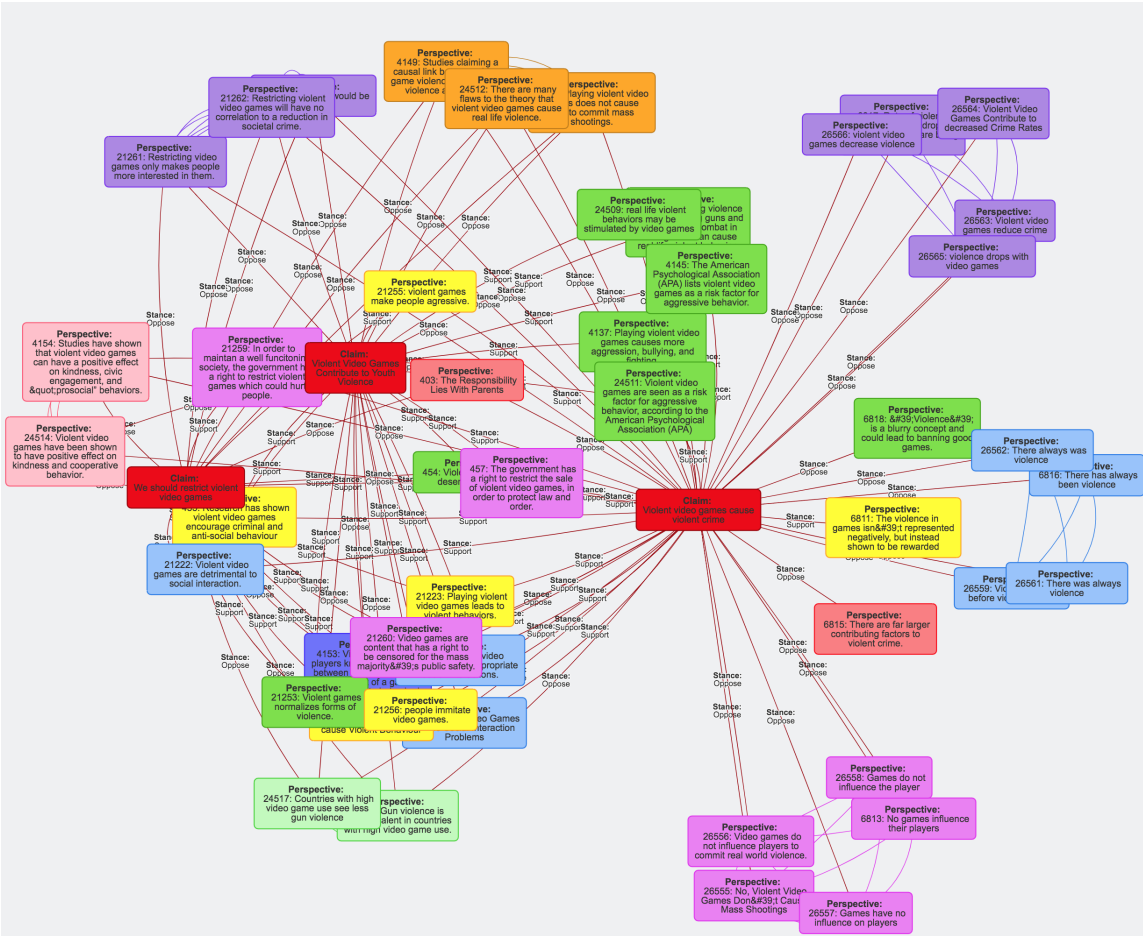

Figure 8: Graph visualization of three related example *claims* (colored in red) in our dataset with their *perspectives*. Each edge indicates a supporting/opposing relation between a perspective and a claim.

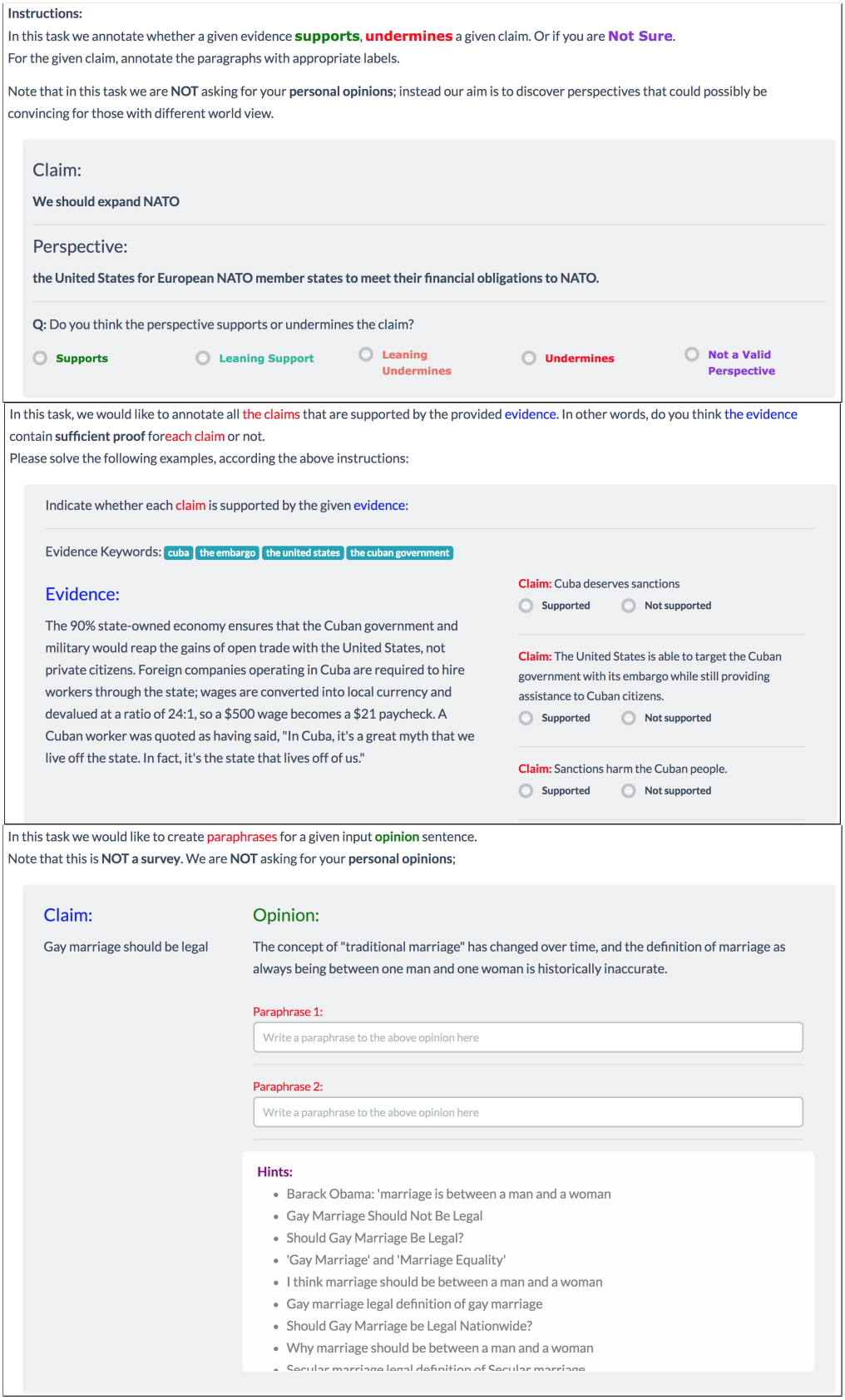

Figure 9: Interfaces shown to the human annotators. Top: the interface for verification of perspectives (step 2a). Middle: the interface for annotation of evidences (step 3a). Bottom: the interface for generation of perspective paraphrases (step 2b).

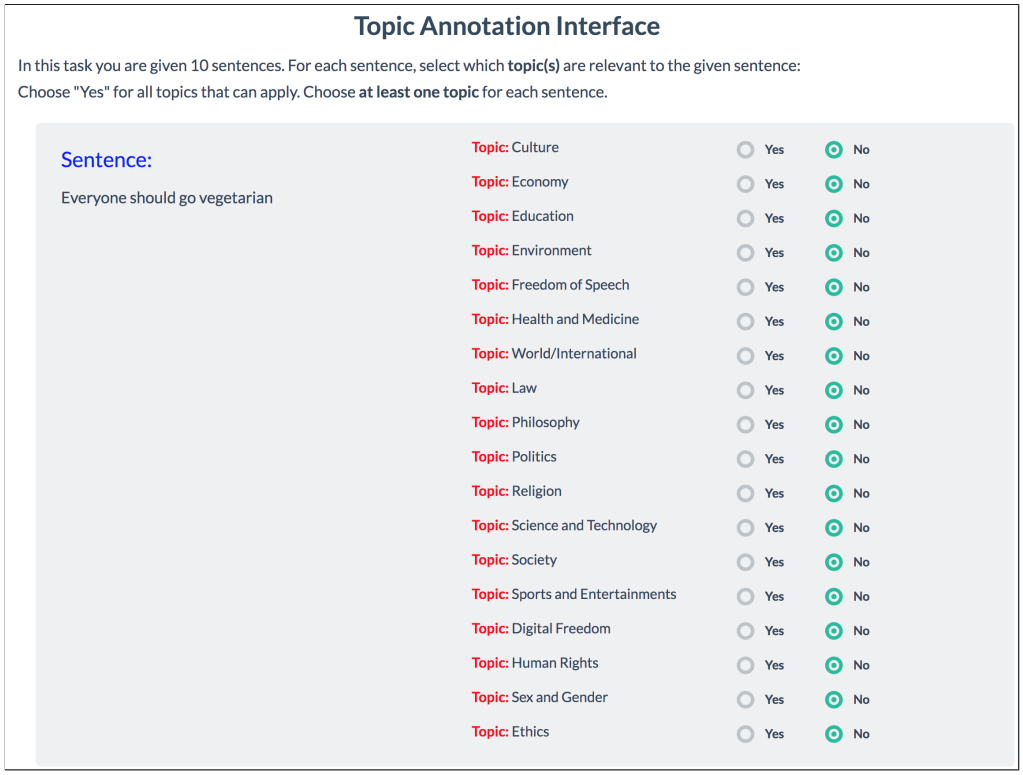

Figure 10: Annotation interface used for topic of claims (Section 4.2)