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.

1 Introduction

Understanding most nontrivial claims requires insights from various perspectives. 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 perspectives 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...


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 \( e \models p|c \).

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 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.
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).

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 ∼ 1k claims, ∼ 8k perspectives and ∼ 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 $\rho$ (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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Stance Classification</th>
<th>Evidence Verification</th>
<th>Human Verified</th>
<th>Open Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSPECTRUM (this work)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FEVER (Thorne et al., 2018)</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(Wachsmuth et al., 2017)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LIAR (Wang, 2017)</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>(Vlachos and Riedel, 2014)</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(Hasan and Ng, 2014)</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>

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

\(^2\)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 ~ 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.

**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 ~ 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

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

Table 2: A summary of PERSPECTRUM statistics
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 sub-set 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 common-sense, 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: \(\{[p_i]\}\).

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 \(U^P\) and \(U^E\), 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 \(U^P\) and one with the pool of evidences \(U^E\). 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 \(U^P\) and evidences \(U^E\).

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

\[
\text{Pre}(c) = \frac{\sum_{p \in \hat{P}(c)} 1\{\exists p, s.t. \hat{p} \in [p]\}}{|\hat{P}(c)|}
\]

and

\[
\text{Rec}(c) = \frac{\sum_{p \in \hat{P}(c)} 1\{\exists p, s.t. \hat{p} \in [p]\}}{|P(c)|}
\]

respectively. To calculate dataset metrics, the aforementioned per-claim metrics are averaged across all the claims in the test set.
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: $\exists \tilde{p} \text{ s.t. } \tilde{p} \in P(c) \land p_1, p_2 \in [\tilde{p}]$. 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 $U_e$. 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 $U_e$. 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 common-sense 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 ~36% over the IR baseline. Meanwhile, this system is behind human performance by a margin of ~20%.

5.3.4 Extraction of supporting evidence (T4)

We evaluate the systems on the extraction of items from the pool of evidences $U^e$, given a claim-perspective pair. To measure the performance of the IR system working with the index containing $U^e$ 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.

### Table 3: Quality of different baselines on different sub-tasks (Section 5).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Target set</th>
<th>System</th>
<th>Pre.</th>
<th>Rec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1: Perspective relevance</td>
<td>$U^p$</td>
<td>IR</td>
<td>46.8</td>
<td>34.9</td>
<td>40.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IR + BERT</td>
<td>47.3</td>
<td>54.8</td>
<td><strong>50.8</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IR + Human</td>
<td>63.8</td>
<td>83.8</td>
<td>72.5</td>
</tr>
<tr>
<td>T2: Perspective stance</td>
<td>$P(c)$</td>
<td>Always “supp.”</td>
<td>51.6</td>
<td>100.0</td>
<td>68.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BERT</td>
<td>70.5</td>
<td>71.1</td>
<td><strong>70.8</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Human</td>
<td>91.3</td>
<td>90.6</td>
<td>90.9</td>
</tr>
<tr>
<td>T3: Perspective equivalence</td>
<td>$P(c)^2$</td>
<td>Always “equiv.”</td>
<td>100.0</td>
<td>11.9</td>
<td>21.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Always “equiv.”</td>
<td>20.3</td>
<td>100.0</td>
<td>33.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IR</td>
<td>36.5</td>
<td>36.5</td>
<td>36.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BERT</td>
<td>85.3</td>
<td>50.8</td>
<td><strong>63.7</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Human</td>
<td>87.5</td>
<td>80.2</td>
<td>83.7</td>
</tr>
<tr>
<td>T4: Evidence extraction</td>
<td>$U^e$</td>
<td>IR</td>
<td>42.2</td>
<td>52.5</td>
<td>46.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IR + BERT</td>
<td>49.7</td>
<td>46.3</td>
<td><strong>55.7</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IR + Human</td>
<td>70.8</td>
<td>53.1</td>
<td>60.7</td>
</tr>
<tr>
<td>T5: Overall</td>
<td>$U^p, U^e$</td>
<td>IR</td>
<td>-</td>
<td>-</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IR + BERT</td>
<td>-</td>
<td>-</td>
<td><strong>17.5</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IR + Human</td>
<td>-</td>
<td>-</td>
<td><strong>40.0</strong></td>
</tr>
</tbody>
</table>

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 claims, perspectives, 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


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.


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)).


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.


V.G. Vinod Vydhiswaran, 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).


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

<table>
<thead>
<tr>
<th>Website</th>
<th># of claims</th>
<th># of perspectives</th>
<th># of evidences</th>
</tr>
</thead>
<tbody>
<tr>
<td>debate</td>
<td>561</td>
<td>4136</td>
<td>4133</td>
</tr>
<tr>
<td>procon</td>
<td>50</td>
<td>960</td>
<td>953</td>
</tr>
<tr>
<td>debatewise</td>
<td>395</td>
<td>3039</td>
<td>3036</td>
</tr>
<tr>
<td>total</td>
<td>1006</td>
<td>8135</td>
<td>8122</td>
</tr>
<tr>
<td>after step 2a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>debate</td>
<td>537</td>
<td>2571</td>
<td>–</td>
</tr>
<tr>
<td>procon</td>
<td>49</td>
<td>619</td>
<td>–</td>
</tr>
<tr>
<td>debatewise</td>
<td>361</td>
<td>1462</td>
<td>–</td>
</tr>
<tr>
<td>total</td>
<td>947</td>
<td>4652</td>
<td>–</td>
</tr>
<tr>
<td>step 2b</td>
<td>paraphrases</td>
<td>–</td>
<td>4507</td>
</tr>
<tr>
<td>step 2c</td>
<td>web perspectives</td>
<td>–</td>
<td>2427</td>
</tr>
</tbody>
</table>

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:

- 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 \( \rho \).
- We can write \( \rho \) 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.
Instructions:
In this task we annotate whether a given evidence supports, undermines a given claim. Or if you are Not Sure.
For the given claim, annotate the paragraphs with appropriate labels.
Note that in this task we are NOT asking for your personal opinions; instead our aim is to discover perspectives that could possibly be convincing for those with different world view.

Claim:
We should expand NATO

Perspective:
the United States for European NATO member states to meet their financial obligations to NATO.

Q: Do you think the perspective supports or undermines the claim?
- Supports
- Leaning Support
- Leaning Undermines
- Undermines
- Not a Valid Perspective

In this task, we would like to annotate all the claims that are supported by the provided evidence. In other words, do you think the evidence contain sufficient proof for each claim or not.
Please solve the following examples, according the above Instructions:

Evidence Keywords: Cuba, the embargo, the united states, the cuban government

Evidence:
The 90% state-owned economy ensures that the Cuban government and military would reap the gains of open trade with the United States, not private citizens. Foreign companies operating in Cuba are required to hire workers through the state; wages are converted into local currency and devalued at a ratio of 24:1, so a $500 wage becomes a $21 paycheck. A Cuban worker was quoted as having said, "In Cuba, it’s a great myth that we live off the state. In fact, it’s the state that lives off of us."

Claim: Cuba deserves sanctions
- Supported
- Not supported

Claim: The United States is able to target the Cuban government with its embargo while still providing assistance to Cuban citizens.
- Supported
- Not supported

Claim: Sanctions harm the Cuban people.
- Supported
- Not supported

In this task we would like to create paraphrases for a given input opinion sentence. Note that this is NOT a survey. We are NOT asking for your personal opinions;

Claim: Gay marriage should be legal

Opinion:
The concept of "traditional marriage" has changed over time, and the definition of marriage as always being between one man and one woman is historically inaccurate.

Parphrase 1:
Write a paraphrase to the above opinion here

Parphrase 2:
Write a paraphrase to the above opinion here

Hints:
- Barack Obama: ‘marriage is between a man and a woman
- Gay Marriage Should Not Be Legal
- Should Gay Marriage Be Legal?
- ‘Gay Marriage’ and ‘Marriage Equality’
- I think marriage should be between a man and a woman
- Gay marriage legal definition of gay marriage
- Should Gay Marriage be Legal Nationwide?
- Why marriage should be between a man and a woman
- Search more on local definitions of marriage nationwide.

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)