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

Sihao Chen et al.

Presenter : Yaman Yu

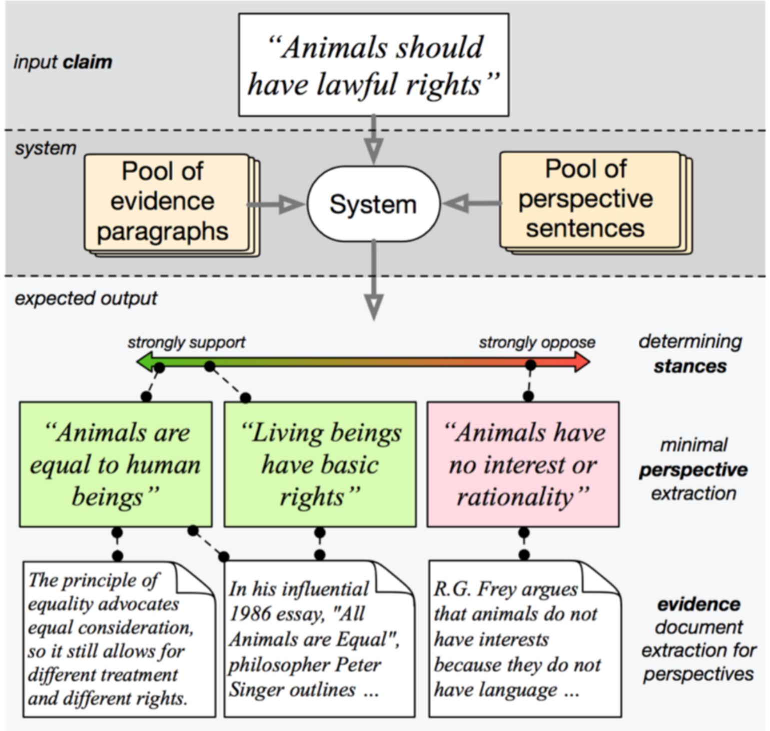
Background

1. Selection bias : We retrieve information by using search engines or recommendation systems. The information we received are based on popularity or our preference.

2. To better understand controversial issues, we need to view them from a diverse and comprehensive set of perspectives with evidence supported.

Example

1. Input a claim -> “*Animal should have lawful rights*”
2. The system is expected to discover various perspectives that are substantiated with evidence and their stance with respect to the claim -> “*Animals are equal to human beings*”


input claim
“Animals should have lawful rights”
system
Pool of evidence paragraphs
System
Pool of perspective sentences
expected output
strongly support
strongly oppose
determining stances
“Animals are equal to human beings”
“Living beings have basic rights”
“Animals have no interest or rationality”
minimal perspective extraction
The principle of equality advocates equal consideration, so it still allows for different treatment and different rights.
In his influential 1986 essay, "All Animals are Equal", philosopher Peter Singer outlines ...
R.G. Frey argues that animals do not have interests because they do not have language ...
evidence document extraction for perspectives

Research Question

1. Identify and formulate NLP tasks for addressing the *substantiated perspectives discovery problem*. **Understand relationship**
   - Between perspectives and claims
   - Nuances of different perspectives
   - Between perspectives and evidence
2. **Build a dataset** for systematic study in the future
3. Develop baseline systems for each sub-task to this problem

Dataset : Perspectrum

Overview: A dataset of *claims, perspectives and evidence paragraphs.* The dataset contains 1k claims, 12k perspectives and 8k evidence paragraphs.

Data source:

- Debate websites as initial seed data: idebate.com, debatewise.com
- Augment with search data
- Using crowdsourcing to increase the quality of the data and clean it from annotation noise

Dataset : Perspectrum

Data construction:

1. Crawl data from debate website -> 1k claims, 8k perspectives, 8k evidences (Significantly noisy and lacks the structure)

2a. Perspective verification: using crowdsourcing (Amazon Mechanical Turk) to hire people verify each perspective is a complete sentence, with a clear stance with respect to the given claim. Also ask them label the stance for each perspective.

2b. Perspective paraphrase: to enrich the ways the perspectives are phrased, they ask people to generate two paraphrases for each of the 15 perspectives

Dataset : Perspectrum

Data construction:

2c. Web perspectives: Using Bing search to query “claim+perspective” to retrieve the 10 most similar sentences then used crowdsourcing annotated

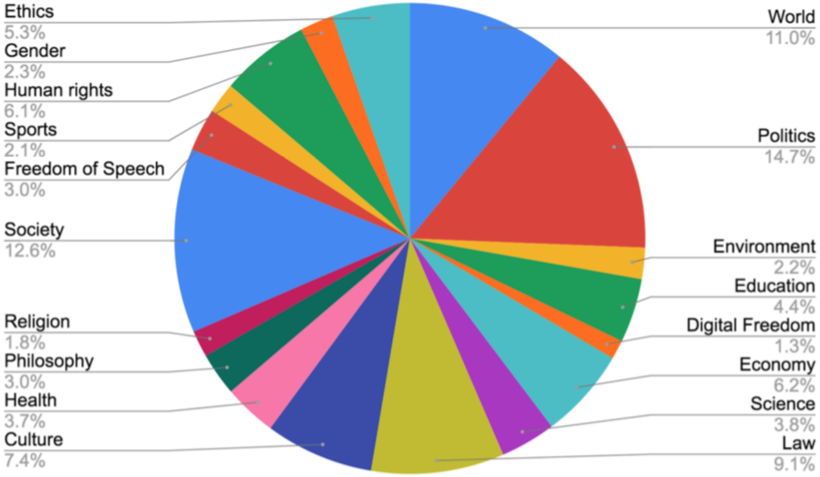
2d. Final perspective trimming: an expert annotator went over all the claims to verify all the equivalent perspectives are clustered together.

3. Evidence verification: for each evidence, they retrieve 8 top relevant perspectives and ask workers on mTurk to annotate whether a given evidence supports a given perspective or not

Dataset : Perspectrum

Data Statistics:

For better understanding the topical breakdown of claims, the paper used crowdsource to annotate the topics of claims. The three topics of Politics, World, and Society have the biggest portions. The general topics of claims are diverse.



NLP Tasks for *substantiated perspectives discovery problem*

1. **Minimal perspective extraction (T1) :** for a input claim, the system is expected to return the collection of mutually disjoint perspectives.
2. **Perspective stance classification (T2) :** for each perspective, the system is expected to label it as *support or oppose* the claim
3. **Perspective equivalence (T3) :** the system is expected to decide whether two given perspectives are equivalent or not
4. **Extraction of supporting evidence (T4) :** for each perspective, the system is expected to return all the evidence from the pool.

Systems Used for NLP Tasks

Information Retrieval (IR): use this system to retrieve a ranked list of best matching perspective/evidence from the corresponding index

Bert: broad range of natural language understanding tasks

Human Performance: use human annotators to measure human performance for each task

Evaluation Metrics for Each NLP Tasks

1. **Minimal perspective extraction (T1) -> Precision & Recall**
2. **Perspective stance classification (T2) -> Precision & Recall**
3. **Perspective equivalence (T3) -> accuracy of two perspectives are in the same cluster for all combinations of perspectives pair**
4. **Extraction of supporting evidence (T4) -> Precision & Recall**
5. **Overall performance -> multiply the disjoint measures in T1, T2 and T4, because T3 has been indirectly measured within T1**

Evaluation Metrics for Each NLP Tasks

1. **Minimal perspective extraction (T1) -> Precision & Recall**

$$\mathrm{Pre}(c) \quad = \quad \frac{\sum_{\hat{p}\in\hat{P}(c)} \mathbf{1}\{\exists p, s.t. \hat{p}\in[\![p]\!]\}}{\left|\hat{P}(c)\right|}$$

$$\mathrm{Rec}(c) \quad = \quad \frac{\sum_{\hat{p}\in\hat{P}(c)} \mathbf{1}\{\exists p, s.t. \hat{p}\in[\![p]\!]\}}{|P(c)|}$$

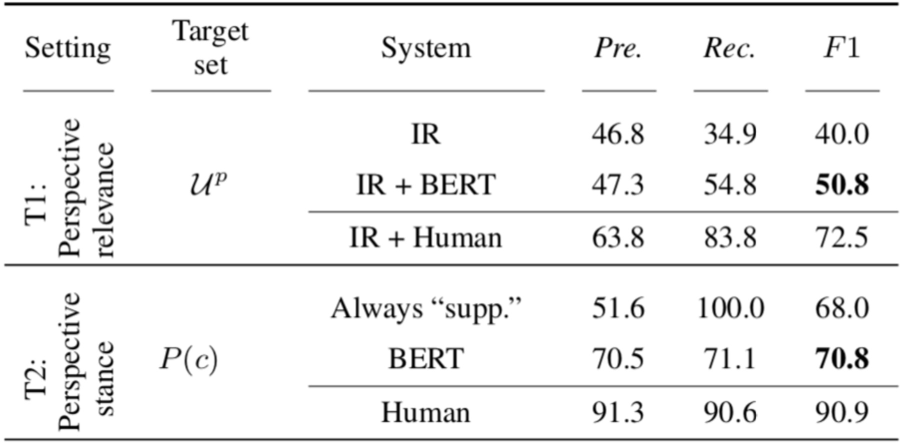
Evaluation Metrics for Each NLP Tasks

4. **Extraction of supporting evidence (T4) -> Precision & Recall**

$$\mathrm{Pre}(p) \;=\; \frac{\left|\hat{E}(p) \cap E(p)\right|}{\left|\hat{E}(p)\right|}$$

$$\mathrm{Rec}(p) \;=\; \frac{\left|\hat{E}(p) \cap E(p)\right|}{|E(p)|}$$

Results of NLP Tasks



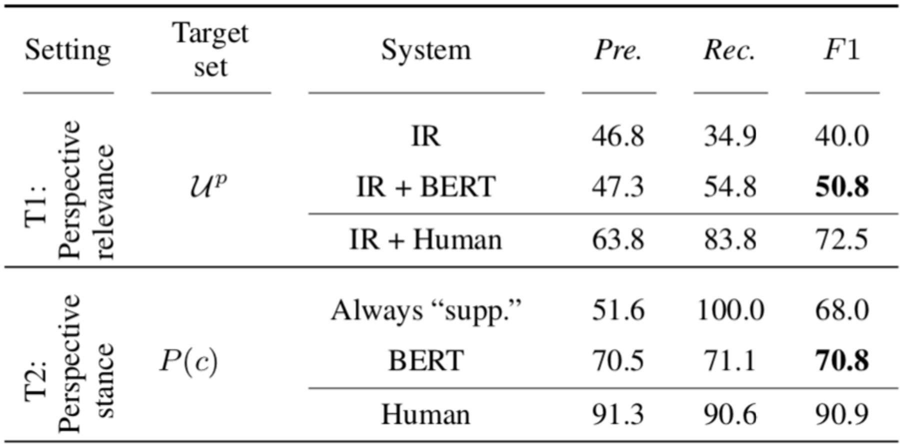
Setting	Target set	System	*Pre.*	*Rec.*	$F1$
T1: Perspective relevance	$\mathcal{U}^p$	IR	46.8	34.9	40.0
		IR + BERT	47.3	54.8	**50.8**
		IR + Human	63.8	83.8	72.5
T2: Perspective stance	$P(c)$	Always "supp."	51.6	100.0	68.0
		BERT	70.5	71.1	**70.8**
		Human	91.3	90.6	90.9

1. Minimal perspective extraction (T1) -> Precision & Recall

Target set: set of perspectives

IR with top-15 candidates yields > 90% recall

Results of NLP Tasks

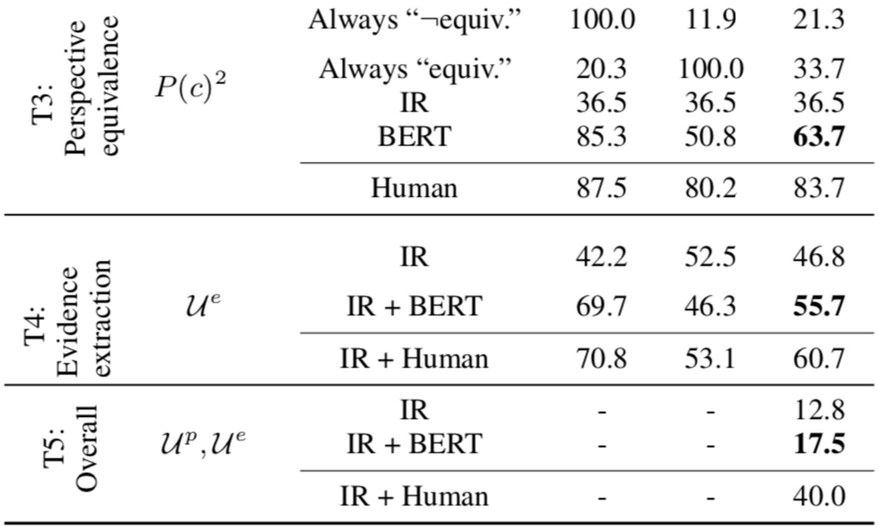


Setting	Target set	System	*Pre.*	*Rec.*	$F1$
T1: Perspective relevance	$\mathcal{U}^p$	IR	46.8	34.9	40.0
		IR + BERT	47.3	54.8	**50.8**
		IR + Human	63.8	83.8	72.5
T2: Perspective stance	$P(c)$	Always “supp.”	51.6	100.0	68.0
		BERT	70.5	71.1	**70.8**
		Human	91.3	90.6	90.9

2. Perspective stance classification (T2) -> Precision & Recall

Target set: set of perspectives related to certain claim

Results of NLP Tasks



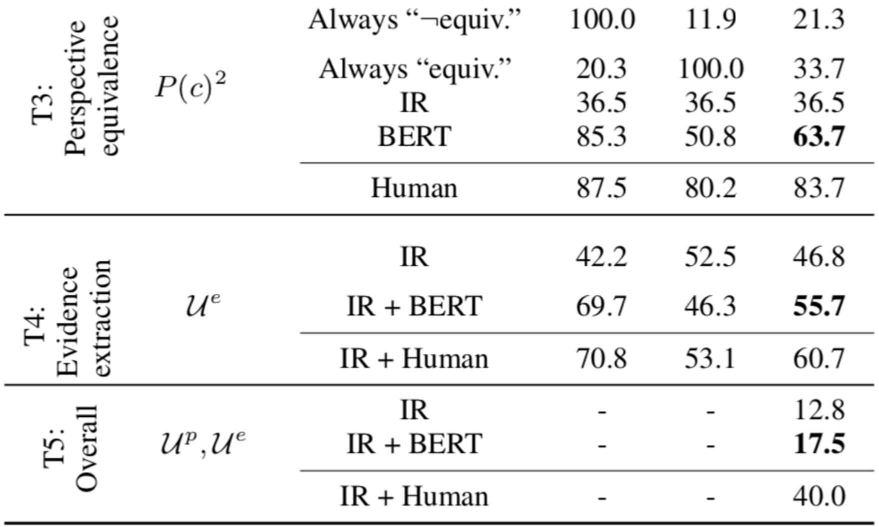
T3: Perspective equivalence	$P(c)^2$	Always "¬equiv."	100.0	11.9	21.3
		Always "equiv."	20.3	100.0	33.7
		IR	36.5	36.5	36.5
		BERT	85.3	50.8	**63.7**
		Human	87.5	80.2	83.7
T4: Evidence extraction	$\mathcal{U}^e$	IR	42.2	52.5	46.8
		IR + BERT	69.7	46.3	**55.7**
		IR + Human	70.8	53.1	60.7
T5: Overall	$\mathcal{U}^p$,$\mathcal{U}^e$	IR	-	-	12.8
		IR + BERT	-	-	**17.5**
		IR + Human	-	-	40.0

3. Perspective equivalence (T3)

-> accuracy of two perspectives are in the same cluster for all combinations of perspectives pair

Target set: all combinations of set of perspectives related to certain claim

Results of NLP Tasks



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

4. Extraction of supporting evidence (T4) -> Precision & Recall

Target set: set of evidence

IR with top-60 candidates yields > 85% recall

Conclusion

1. This work define the problem of substantiated perspective discovery and NLP tasks related to this problem
2. The paper build the dataset by combing online resources, web data and crowdsourcing to bring more attention to this problem
3. They also build baseline and evaluation method for each NLP task

Future Work

The paper assumed that the input claims are valid and contradictory, which is not always true. So one of the future work could be develop mechanism to recognize valid argumentative structures.