A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics

Ehud Aharoni[1 2], Anatoly Polnarov[2 3], Tamar Lavee[1 4], Daniel Herchcovich[1], Ran Levy[1], Ruty Rinott[1], Dan Gutfreund[1], Noam Slonim[1 5]

Abstract

We describe a novel and unique argumentative structure dataset. This corpus consists of data extracted from hundreds of Wikipedia articles using a meticulously monitored manual annotation process. The result is 2,683 argument elements, collected in the context of 33 controversial topics, organized under a simple claim-evidence structure. The obtained data are publicly available for academic research.

1 Introduction

One major obstacle in developing automatic argumentation mining techniques is the scarcity of relevant high quality annotated data. Here, we describe a novel and unique benchmark data that relies on a simple argument model and elaborates on the associated annotation process. Most importantly, the argumentative elements were gathered in the context of pre-defined controversial topics, which distinguishes our work from other previous related corpora.[6] Two recent works that are currently under review [Rinott et al, Levy et al] have reported first results over different subsets of this data, which is now publically available for academic research upon request. We believe that this novel corpus should be of practical importance to many researches, and in particular to the emerging community of argumentation mining.

Unlike the classical Toulmin model (Freeley and Steinberg 2008) we considered a simple and robust argument structure comprising only two components – *claim* and associated supporting *evidence*. The argumentative structures were carefully annotated under a pre-defined *topic*, introduced as a debate motion. As the collected data covers a diverse set of 33 motions, we expect it could be used to develop generic tools for automatic detection and construction of argumentative structures in the context of new topics.

2 Data Model

We defined and used the following concepts:
Topic – a short phrase that defines the subject of interest and the sentiment towards it. **Context Dependent Claim (CDC)** – a general concise statement that directly supports or contests the Topic. **Context Dependent Evidence (CDE)** – a text segment that directly supports a CDC in the context of a given Topic. Examples for these concepts are given in Section 6.

Since one can support a claim using different types of evidence (Rieke et al 2012, Seech 2008), we defined and considered three CDE types: **Study**: Results of a quantitative analysis of data given as numbers or as conclusions. **Expert**: Testimony by a person / group / committee / organization with some known expertise in or authority

[1] IBM Haifa Research Lab, Haifa, Israel.

[2] These authors contributed equally to this manuscript.

[3] Hebrew University, Israel.

[4] Present affiliation – Yahoo! Labs.

[5] Corresponding author, at noams@il.ibm.com.

[6] E.g., AraucariaDB (Reed 2005, Moens et al 2007) and Vaccine/Injury Project (V/IP) Corpus (Ashley and Walker 2013).

on the topic. **Anecdotal:** a description of specific event(s)/instance(s) or concrete example(s).[7]

3 Labeling Challenges and Approach

The main challenge we faced in collecting the annotated data was the inherently elusive nature of concepts such as "claim" and "evidence." To address that we formulated two sets of criteria for CDC and CDE, respectively, and relied on a team of carefully trained in-house labelers whose consequent work was closely monitored. To further enhance the quality of the collected data we employed a two-staged labeling approach. First, a team of five labelers worked independently on the same text and prepared the initial set of candidate CDCs or candidate CDEs. Next, a typically different team of five labelers independently crosschecked the joint list of the detected candidates, each of whom was either confirmed or rejected. Candidates confirmed by at least three labelers were included in the corpus.

4 Labeling Guidelines

The labeling guidelines defined the concepts of Topic, CDC, CDE, and CDE types, along with relevant examples. According to these guidelines, given a Topic, a text fragment should be labeled as a CDC if and only if it complies with all of the following five CDC criteria: **Strength**: Strong content that directly supports or contests the provided Topic. **Generality**: General content that deals with a relatively broad idea. **Phrasing**: Well phrased, or requires at most a single and minor "allowed" change.[8] **Keeping text spirit**: Keeps the spirit of the original text. **Topic unity**: Deals with one, or at most two related topics.

Given a Topic and a CDC, a text fragment should be labeled as a CDE if and only if it complies with all of the following four CDE criteria: **Strength**: Strong content that directly supports the CDC's main point in the context of the Topic. **Phrasing**: Reasonably well phrased, and easily understandable given the CDC and the Topic. **Keeping text spirit**: Keeps the spirit of the original text. **Unity**: Cannot be naturally divided into individual CDEs, nor joined to previous and/or successive text to generate clearly stronger CDE.

5 Labeling Details

The labeling process was carried out in the GATE environment (https://gate.ac.uk/). The 33 Topics were selected at random from the debate motions at http://idebate.org/ database. The labeling process was divided into five stages:

Search: Given a Topic, five labelers were asked to independently search English Wikipedia[9] for articles with promising content. Specifically, an article should have been deemed appropriate for labeling if and only if the labeler thought it contained at least three CDCs.

Claim Detection: At this stage, five labelers independently detected candidate CDCs supporting or contesting the Topic within each article suggested by the Search team.

Claim Confirmation: At this stage, five labelers independently cross-examined the candidate CDCs suggested at the Claim Detection stage, aiming o confirm a candidate and its sentiment as to the given Topic, or reject it by referring to one of the five CDC Criteria. The resulting list of confirmed CDCs comprised the candidate CDCs confirmed by at least three labelers.

Evidence Detection: At this stage, five labelers independently detected candidate CDEs support-

[7] The full guidelines include additional, typically less frequent types of CDE. These will be discussed in the full version of this paper.

[8] For example - pronoun resolution. The enclosed data set contains the corrected version as well, as proposed by the labelers.

[9] We considered a Wikipedia dump from April 2012.

ing a confirmed CDC in the context of the given Topic. At this point, the search for CDEs was done only in the same article where the corresponding CDC was found.

Evidence Confirmation: At this stage, five labelers were asked either to confirm each candidate CDE, and classify it under one or more of the CDE Types, or reject it by referring to one of the four CDE Criteria. The resulting list of the confirmed CDEs comprised the candidate CDEs confirmed by at least three labelers.

Labelers training and feedback: Before joining actual labeling tasks, novice labelers were assigned with several completed tasks and were expected to show a reasonable degree of agreement with a consensus solution prepared in advance by the project administrators. In addition, the results of each Claim Confirmation task were examined by one or two of us (AP and NS) to ensure the conformity to the guidelines. In case crude mistakes were spotted, the corresponding labeler was requested to revise and resubmit. Due to the large numbers of CDE candidates, it was impractical to rely on such a rigorous monitoring process in Evidence Confirmation. Instead, Evidence Consensus Solutions were created for selected articles by several experienced labelers, who first solved the tasks independently and then reached consensus in a joint meeting. Afterwards, the tasks were assigned to the rest of the labelers. Their results were juxtaposed with the Consensus Solutions, and on the basis of this comparison individual feedback reports were drafted and sent to the team members. Each labeler received such a report on an approximately weekly basis.

6 Data Summary

For 33 debate motions, a total of 586 Wikipedia articles were labeled. The labeling process resulted with 1,392 CDCs, distributed across 321 articles. For 12 debate motions, for which 350 distinct CDCs were confirmed across 104 articles, we further completed the CDE labeling, ending up with a total of 1,291 confirmed CDEs – 431 of type Study, 516 of type Expert, and 529 of type Anecdotal. Note, some CDEs were associated with more than one type. For example, 118 CDEs were assigned type Study and type Expert.

In Tables 1 and 2 we present several examples of CDCs and CDEs gathered under the Topics we worked with. Further, we present several inacceptable candidates in order to exemplify some of the subtleties of the performed work.

Topic	The sale of violent video games to minors should be banned
Pro- CDC	*Violent video games can increase children's aggression*
Pro- CDC	*Video game publishers unethically train children in the use of weapons* Note, that a valid CDC is not necessarily factual.
Con- CDC	*Violent games affect children positively*
Invalid CDC 1	*Video game addiction is excessive or compulsive use of computer and video games that interferes with daily life.* This statement defines a concept relevant to the Topic, not a relevant claim.
Invalid CDC 2	*Violent TV shows just mirror the violence that goes on in the real world.* This claim is not relevant enough to Topic.
Invalid CDC 3	*Violent video games should not be sold to children.* This candidate simply repeats the Topic, and thus is not considered a valid CDC.

Table 1: Examples of CDCs and invalid CDCs.

Topic 1	The sale of violent video games to minors should be banned
CDC	*Violent video games increase youth violence*
Study CDE	*The most recent large scale meta-anlysis-- examining 130 studies with over 130,000 subjects worldwide -- concluded that exposure to violent video games causes both short term and long term aggression in players*
Anecdotal CDE	*In April 2000, a 16-year-old teenager murdered his father, mother and sister pro-*

	claiming that he was on an "avenging mission" for the main character of the video game Final Fantasy VIII
Invalid CDE 1	*Studies have been conducted to prove the effects of violent video games on children and adolescents.* Invalid, since the studies' conclusion is not mentioned.
Invalid CDE 2	*While most experts reject any link between video games content and real-life violence, some media scholars argue that the connection exists.* Invalid, because it includes information that contests the CDC.
Topic 2	The use of performance enhancing drugs in sports should be permitted
CDC	*Drug abuse can be harmful to one's health and even deadly.*
Expert CDE	*According to some nurse practitioners, stopping substance abuse can reduce the risk of dying early and also reduce some health risks like heart disease, lung disease, and strokes*
Invalid CDE	*Suicide is very common in adolescent alcohol abusers, with 1 in 4 suicides in adolescents being related to alcohol abuse.* Although the candidate CDE does support the CDC, the notion of adolescent alcohol abusers is irrelevant to the Topic. Therefore, the candidate is invalid.

Table 2: Examples of CDEs and invalid CDEs.

7 Agreement and Recall Results

To evaluate the labelers' agreement we used Cohen's kappa coefficient (Landis and Koch 1977). The average measure was calculated over all labeler's pairs, for each pair taking those articles on which the corresponding labelers worked together and omitting labeler pairs which labeled together less than 100 CDCs/CDEs. The obtained average kappa was 0.39 and 0.4 in the Claim confirmation and Evidence confirmation stages, respectively, which we consider satisfactory given the subtlety of the concepts involved.

We further employed a simple method to obtain a rough estimate of the recall at the detection stages. For CDCs (and similarly for CDEs), let n be the number of CDCs detected and confirmed in a given article, and x be the unknown total number of CDCs in this article. Assuming each labeler detects a ratio p_i of x, and taking a strong assumption of independence between the labelers, we get:

$$x \prod_i (1 - p_i) = x - n,$$

We estimated p_i from the observed data, and computed x for each article. We were then able to compute the estimated recall per motion, ending up with the estimated average recall of 90.6% and 90.0% for CDCs and CDEs, respectively.

8 Future Work and Conclusion

There are several natural ways to proceed further. First, a considerable increase in the quantity of gathered CDE data can be achieved by expanding the search scope beyond the article in which the CDC is found. Second, the argument model can be enhanced – for example, to include counter-CDE (i.e., evidence that contest the CDC). Third, one may look into ways to add more labeling layers on top of the existing model; for example, distinguishing between factual CDCs, value CDCs, and so forth. Fourth, new topics and new sources besides Wikipedia can be considered.

The data is released and available upon request for academic research. We hope that it will prove useful for different data mining communities, and particularly for various purposes in the field of corpus-based discourse analysis. As noted above, two works exemplifying possible applications of the described dataset are currently under review [Rinott et al, Levy et al].

References

Austin J. Freeley and David L. Steinberg. 2008. *Argumentation and Debate*. Wadsworth, Belmont, California.

Chris Reed. 2005. “Preliminary Results from an Argument Corpus” in *Proceedings of the IX Symposium on Social Communication*, Santiago de Cuba, pp. 576-580.

J. Richard Landis and Gary G. Koch. 1977. “The measurement of observer agreement for categorical data.” *Biometrics* 33:159-174.

Kevid D. Ashley and Vern R. Walker. 2013. “Toward Constructing Evidence-Based Legal Arguments Using Legal Decision Documents and Machine Learning” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (ICAIL ’13)*, Rome, Italy, pp. 176-180.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. “Automatic Detection of Arguments in Legal Texts” in *Proceedings of the International Conference on AI & Law (ICAIL-2007)*, Stanford, CA, pp. 225-230.

Richard D. Rieke, Malcolm O. Sillars and Tarla Rai Peterson. 2012. *Argumentation and Critical Decision Making (8e)*. Prentice Hall, USA.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni and Noam Slonim. “Context Dependent Claim Detection.” Submitted

Ruty Rinott, Lena Dankin, Carloz Alzate, Ehud Aharoni and Noam Slonim. "Show Me Your Evidence – an Automatic Method for Context Dependent Evidence Detection.” Submitted.

Zachary Seech. 2008. *Writing Philosophy Papers (5th edition)*. Wadsworth, Cengage Learning, Belmont, California.