The 2016 AAAI Fall Symposium Series:
Cognitive Assistance in Government and Public Sector Applications
Technical Report FS-16-02

Automatic Arguments Construction – From Search Engine to Research Engine

Dan Gutfreund, Yoav Katz, Noam Slonim
IBM Debating Technologies Team
IBM Research
ranita@il.ibm.com

Abstract

While discussing a concrete controversial topic, most humans will find it challenging to swiftly raise a diverse set of convincing and relevant arguments. In this paper we present a system that, given a point of view about a controversial topic, automatically generates arguments supporting and contesting it. This is achieved by breaking the task of automatic argument construction into a pipeline of successive modules, each is responsible for a specific tangible task such as documents retrieval, identifying building blocks of arguments within a document, and analyzing whether these building blocks support or contest the point of view. By providing an interface for humans to interact and intervene at different points in the pipeline, we present an interactive research tool which, for a given topic and a corpus of documents such as Wikipedia or newspaper archive, provides a more comprehensive view and deeper insights than can be obtained using standard search engines.

Introduction

The ability to argue in a persuasive manner is an important aspect of human interaction that naturally arises in various domains such as politics, marketing, law, and health-care. Furthermore, good decision making relies on the quality of the arguments being presented and the process by which they are resolved. Thus, it is not surprising that argumentation has long been a topic of interest in academic research, and different models have been proposed to capture the notion of an argument (Freeley and Steinberg 2008).

A fundamental component which is common to all these models is the concept of *claim* (or *conclusion*). Specifically, at the heart of every argument lies a single claim, which is the assertion the argument aims to prove. The proof of a claim is given by one or more *evidence* (or *premise*). Here we use evidence in a very broad sense, meaning that it can be a study supporting the claim, an expert opinion, a related historic event etc.

Given a concrete topic, or context, most humans will find it challenging to swiftly raise a diverse set of convincing and relevant claims and supporting evidence. In this work we present a system that, given a point of view about a controversial topic, automatically pinpoints relevant claims and evidences in a given corpus, determines their polarity with respect to the given point of view, allows the user to modify various steps, and articulates them per the user's request.

Basic Concepts and Associated Challenges

We define and rely on the following concepts (see Table 1 for examples):

Topic: Short, usually controversial, statement that defines a point of view on the subject of interest.

Context Dependent Claim (CDC): General and concise statement that directly supports or contests the given Topic.

Context Dependent Evidence (CDE): A text segment that directly supports a claim in the context of a given topic.

Given these definitions, as well as a few more detailed criteria to reduce the variability in the manually labeled data, human labelers were asked to detect CDCs and CDEs for a diverse set of Topics in relevant articles taken from various sources such as Wikipedia. The collected data were used to train and assess the performance of the statistical models that underlie our system. These data are freely available for academic research (Aharoni et. al. 2014, https://www.research.ibm.com/haifa/dept/vst/mlta_data.shtml). The examples and statistics below are based on these data.

The distinction between a CDC and other related texts can be quite subtle, as illustrated in Table 1.

For example, automatically distinguishing a CDC like S1 from a statement that simply defines a relevant concept like S4, from a claim which is not relevant enough to the given Topic like S5, from a statement like S6 that merely repeats the given Topic in different words, or from a statement that represents a relevant claim which is not general enough like S7, is clearly challenging. Further, CDCs can be of different flavors, ranging from factual assertions like S1 to statements that are more a matter of opinion (Pang and Lee 2008) like S2, adding to the complexity of the task. Moreover, our data suggest that even if one focuses on Wikipedia articles that are highly relevant to the given Topic, only ≈2% of their sentences include CDCs.

Topic	The sale of violent video games to minors should be banned
(Pro) CDC	*S1: Violent video games can increase children's aggression*
(Pro) CDC	*S2: Video game publishers unethically train children in the use of weapons* Note, that a valid CDC is not necessarily factual.
(Con) CDC	*S3: Violent games affect children positively*
Invalid CDC 1	*S4: 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	*S5: Violent TV shows just mirror the violence that goes on in the real world.* This statement is not relevant enough to the Topic.
Invalid CDC 3	*S6: Violent video games should not be sold to children.* This statement simply repeats the Topic, and thus is not considered a valid CDC.
Invalid CDC 4	*S7: "Doom" has been blamed for nationally covered school shooting*. This statement is not general enough to represent a CDC, as it focuses on a specific single video game (although it can potentially be used as a CDE).

Table 1. Examples of CDCs and invalid CDCs.

Furthermore, since CDCs are by definition concise statements, they typically do not span entire Wikipedia sentences but rather sub-sentences. This is illustrated in Table 2. There are many optional boundaries to consider when trying to identify the exact boundaries of a CDC within a typical Wikipedia sentence. This task further complicates the CDC detection problem. Thus, we are faced with a large number of candidate CDCs, of which only a tiny fraction represents positive examples. Furthermore, the distinction between positive and negative examples can be quite subtle. Finally, even if a correct CDC is detected, automatically determining its Pro/Con polarity with respect to the Topic poses additional unique challenges.

Because violence in video games is interactive and not passive, critics such as Dave Grossman and Jack Thompson argue that ***violence in games hardens children to unethical acts****, calling first-person shooter games ``murder simulators'', although no conclusive evidence has supported this belief.*

Table 2. A CDC is often only a small part of a single Wikipedia sentence - e.g., the part marked in bold in this example.

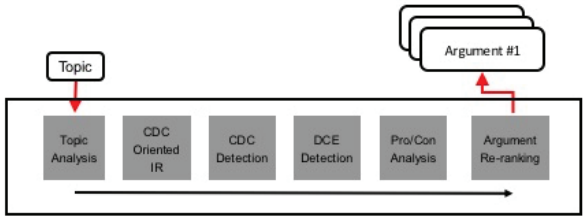

Figure 1. High level architecture of the demonstrated system.

Similar challenges arise in the treatment of CDEs. In addition to these challenges, it is well recognized that claims can be supported by different types of evidence (Reike and Sillars 2001, Seech 2008). Thus, it is expected that in a given free text data, different evidence types will be associated with different statistical signatures. Correspondingly we develop an individual classification approach to three common evidence types (Rinott et. al. 2015):

Study – Results of a quantitative analysis given as numbers or as conclusions.

Expert – Testimony of a person/group/committee/organization with some known expertise / authority on the topic

Anecdotal – Description of specific event(s), instance(s) or concrete example(s).

High Level Architecture

The system relies on a cascade of engines, depicted in Figure 1. In general, these engines rely on various Information Retrieval (IR), Natural Languuage Processing (NLP), and Machine Learning (ML) techniques, as well as different resources and lexicons like WordNet (Miller 1995).

Given a Topic and a corpus of articles such as Wikipedia or a newspaper archive, the Topic Analysis engine starts with initial semantic analysis of the Topic, aiming to identify the main concepts mentioned in this Topic and the sentiment towards each concept. Next, the CDC Oriented Article Retrieval engine employs IR and opinion mining techniques in order to retrieve articles that with high probability contain CDCs. A detailed description of a preliminary version of this engine can be found in (Roitman et. al. 2016). Next, the CDC Detection engine relies on a combination of NLP and ML techniques to zoom-in within the retrieved articles and detect candidate CDCs. A detailed description of this engine can be found in (Levy et. al. 2014). Next the CDE Detection engine relies on a combination of NLP and ML techniques to identify evidences of various types within the retrieved articles and to associate them with claims that were detected in the previous stage. A detailed description of this engine can be found in (Rinott et. al. 2015). Next the Pro/Con engine aims to automatically determine the polarity of the candidate CDC and CDE with respect to the given Topic by analyzing and contrasting the sentiment towards key concepts mentioned in the Topic and within the candidate CDC/CDE (Bar-Haim et. al. 2016). Finally, the Argument re-ranking engine aims to improve the precision of the generated output, based on the results collected thus far; e.g., using a simple rule-based approach, we remove candidate CDCs for which the predicted Pro/Con polarity has low confidence.

Research Engine

The pipeline described in the previous section mimics at a high level the steps that humans may take when conducting a research on a given controversial topic. First, search for the relevant articles; then, identify within the articles the main claims with respect to the topic; finally, look for evidences of various types to support each of these claims. A natural idea is to provide an interface for humans to interact with the system in the different stages of the pipeline, thus allowing them to guide the analysis that the system performs and correct errors along the way (this can later be used to retrain and improve the system). With such a system at hand, a user can swiftly generate a list of high quality arguments according to criteria of his choice, whose building blocks (claims and evidences) are taken from a very large and diverse corpora of articles. Note that the system can generate arguments that have never been stated before, by say gluing together a claim from article A with expert evidence from article B and study evidence from article C. In addition, by learning the patterns of previously seen claims, the system can generate new claims altogether (Bilu and Slonim 2016). We coin this tool *Research Engine.*

Clearly, a research engine's capabilities go a long way beyond what is achievable with existing search engines. It is instructive to point out the differences between the two. A search engine receives a simple query, typically key words, and very quickly returns a list of relevant articles. This is typically achieved by first performing an extensive pre-processing computation to create an index. This heavy computation is query (or context) independent. Then, when the query is received, the search engine performs a quick *query dependent* computation over the index to obtain the results. A research engine on the other hand, may receive a sentence in natural language and generate a list of arguments by identifying claims and evidences in different articles from which it aims to compose whole arguments. Similar to search engines, it relies on indexing for the article retrieval stage. However, once it obtains the articles, it performs a heavy *query dependent* computation to zoom in and detect the argumentative units of text within these articles, identify their polarity, inter-relations, and so forth. Thus a research engine will typically follow a much more demanding process per query, compared to the classical search engine, but correspondingly the results will be of a much higher quality, specificity, and depth. In this tradeoff between the processing requirements of a query vs. the output quality, the light computation / shallow results regime (namely search engines) has received a lot of attention. In this work we shift the focus to the less explored regime of heavy computation / in-depth results.

Summary

In this paper we describe a system that given a controversial topic automatically generates arguments that either support or contest the topic by sifting through massive textual corpora and identifying relevant claims and evidences. By providing an interface that allows humans to interact with the system at various stages, the presented research engine allows humans to conduct research for a controversial topic at much greater pace. We believe that this tool will pave the way to a new generation of cognitive assistants that can swiftly analyze huge corpora, extract the most relevant information (and not just a list of articles) and interact with the user in a way that allows the user to guide the research process toward the most suitable outcome for her needs.

References

Freeley, A. J.; and Steinberg D. L, 2008. *Argumentation and Debate*, Wadsworth.

Aharoni E.; Polnarov A.; Lavee T.; Hershcovich D.; Levy R.; Rinott R.; Gutfreund D.; and Slonim N. 2014. A Benchmark Dataset for Automatic Detection of Claims and Evidence in the

Context of Controversial Topics, in *Proceedings of First Workshop on Argumentation and Computation, 52nd ACL.*

Pang B.; and Lee L. 2008. Opinion mining and sentiment analysis, in *Foundations and Trends in Information Retrieval*, Vol. 2, pp. 1-135.

Reike R. D.; and Sillars M. O. 2001. *Argumentation and critical decision making*, Longman.

Seech Z. 2008. *Writing philosophy papers*, Cengage Learning.

Rinott R.; Dankin L.; Alzate C. P.; Khapra M. M.; Aharoni E.; and Slonim N. 2015. Show me your evidence – an automatic method for context dependent evidence detection, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*: 440-450.

Miller G. A. 1995. WordNet: A Lexical Database for English, in *Communications of the ACM*, 38: 29-41.

Roitman H.; Hummel S.; Rabinovich E.; Sznajder B.; Slonim N; Aharoni E. 2016. On the retrieval of Wikipedia articles containing claims on controversial topics, *WWW (Companion Volume):* 991-996.

Levy R.; Bilu Y.; Hershcovich D.; Aharoni E.; and Slonim N. 2014. Context Dependent Claim Detection, In *Proceedings of the* 25th *International Conference on Computational Linguistics:* 1489-1500.

Bar-Haim R.; Bhattacharya I.; Dinuzzo F.; Saha A.; and Slonim N. 2016. Stance Classification of Context-Dependent Claims, Submitted.

Bilu Y.; and Slonim N. 2016. Claim Synthesis via Predicate Recycling, in *Proceedings of The* 54th *Annual meeting of the Association for Computational Linguistics.*