Annotation of Argument Components in Political Debates Data

Shohreh Haddadan, Elena Cabrio, Serena Villata
University of Luxembourg, Université Côte d'Azur, Université Côte d'Azur
shohreh.haddadan@uni.lu, elena.cabrio@unice.fr , villata@i3s.unice.fr

Abstract

In this paper, we present the annotation guidelines we defined for annotating arguments in political debates. In our guidelines, we consider each argument as being composed of a claim and one or more premises. The annotation process has started with defining the guidelines for three annotators containing examples from the data, and continued as cyclic process of evaluation and revision on the annotation to resolve the ambiguities in the guidelines. In this paper, we briefly discuss the resulting annotated dataset and give some examples of the annotation scheme. The quality of the annotated dataset is assessed by computing inter-annotator agreement using Krippendorf's α coefficient on a portion of the dataset.

Keywords: Argumentation Mining, Argument Components, Annotation, Inter-annotator Agreement

1. Introduction

The ability to make persuasive arguments is a crucial characteristic for politicians in order to convince the public to vote for them. Moreover, social scientists and historians are interested in following the logical process of arguments proposed by politicians. For instance, in the book "Lincoln, Doglas and Slavery" (Zarefsky, 1993), the author analyses the two presidential candidate's arguments and their reasoning strategies toward the specific issue of slavery. Thus, social scientists and historians would highly benefit from a tool able to assist them in detecting the arguments from natural language documents.

The rising field of Argument(ation) Mining (Lippi and Torroni, 2016b) aims at establishing the foundations of formulating arguments and their reasoning process by extracting argument components and the relations between them from various types of textual resources. The argumentation mining pipeline, more precisely, consists of several stages which apply Natural Language Processing (NLP) and Machine Learning (ML) methods. These stages include the detection of argumentative vs non-argumentative utterances, the classification of the argumentative utterances with respect to the two classes of argument components (i.e., premises and claims), and the prediction of the kind of relation holding between these components. The output is a structure where each argument is connected to the others, providing in this way an overall view of the argumentation. Supervised machine learning methods need annotated data in the training phase. Hence, the need for annotated data as an input to the argumentation mining pipeline is gaining more significance.

In this work, we first introduce the raw data used for building this dataset in Section 2. Subsequently, in Section 3 we explain the scheme for argument component annotation we defined by providing examples from the dataset. Then, we explain the annotation process and the training process of annotators in Section 4. Moreover, we discuss the challenges which we confronted during the annotation process relevant to the choice of dataset in Section 5. As a result for this study we give an estimation of the quality of the annotated dataset by computing the inter-annotator agreement over a portion of the annotated dataset in Section 6. Section 7 concisely discusses previous work in argument mining from political data.

2. Dataset

The dataset used for this study is taken from the Commission on Presidential Debates (CPD) website which is an independent nonprofit corporation sponsoring U.S. presidential and vice-presidential debates. This dataset contains the transcripts of debates that have been publicly broadcast[1]. The first presidential debate ever held on television was the debate between Kennedy and Nixon in 1960. Despite the enormous audiences for the Kennedy-Nixon encounters, 16 years went by before the next series of debates. Thus, no debate transcript exists in the dataset for 1964, 1968 and 1972. The dataset therefore includes 12 sets of debates from 1960 to 2016 presidential election debates.

The dataset consists of 41 different transcripts in 12 years. Table 1 shows a summary of the sum and average number of turns of speech, sentences and tokens over the debate transcripts of the whole corpus.

	Turns of Speech	Sentences	Tokens
SUM	6,907	36,988	678,291
AVERAGE	160.6	860.186	15,774.21

Table 1: Sum and average of turns of speech, sentences and tokens in the dataset.

3. Annotation Scheme

In this study, we aim at proposing an annotation scheme for the annotation of argument components in political debates. As mentioned in Section 1, arguments consist of two major components, in this section we will provide the definitions and some examples from the dataset for these argument components which represent the key element of the annotation scheme for our dataset. The argument components annotated for this dataset are claims and premises.

According to Toulmin's model of reasoning, the basic triad of an argument consists of three main components: claim,

[1] http://www.debates.org/

data and warrant (Toulmin, 2003). Claims are the basic component of arguments. Claims are made so that the audience of the argument accept them, they can also be considered as conclusion of the argument. Data is the component of the argument which is provided in order to support the truth of the made claim. Warrants are components which connect the claim to data by certifying whether its reasonable. Warrants are typically implicit and are not stated on the premise that the audience can infer them. Figure 1 illustrates Toulmin's model's basic triad by giving an example from John Kennedy's speech in debate against Richard Nixon in 1960: [2]

“[**In my judgment, the hard money, tight money policy, fiscal policy of this Administration has contributed to the slow-down in our economy**], [*which helped bring the recession of fifty-four*]; [*which made the recession of fifty-eight rather intense*], and [*which has slowed, somewhat, our economic activity in 1960*].”

Our annotation scheme includes the annotation of argument components which are claims and premises (referred to as data in Toulmin's model).

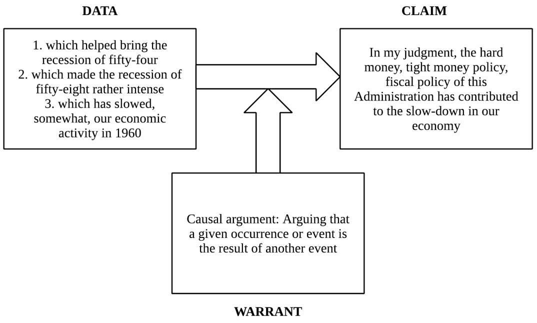
DATA
CLAIM
1. which helped bring the recession of fifty-four
2. which made the recession of fifty-eight rather intense
3. which has slowed, somewhat, our economic activity in 1960
In my judgment, the hard money, tight money policy, fiscal policy of this Administration has contributed to the slow-down in our economy
Causal argument: Arguing that a given occurrence or event is the result of another event
WARRANT

Figure 1: The Toulmin model of argument applied to Kennedy's speech in debate against Nixon 26 September 1960

3.1. Claim

In political debates, a claim may suggest a policy advocated by a party or a candidate to be undertaken which needs to be justified in order to be accepted by the audience. Claims may also be made to take a stance towards a certain policy. They might state an opinion or belief or judgment towards a specific issue.

In order to help annotators in finding claims we have suggested some indicator phrases which were commonly used by the candidates while making claims, such as: I believe, in my opinion, I am in favor of, I think ,etc.

Political debates are rifled with all types of various claims, in order to make the definitions clear some examples are provided from the dataset in the following.

1. “[**I feel that another effort should be made by a new Administration in January of 1961, to renew negotiations with the Soviet Union**]”.[3]
2. “[**George Bush, who I think is one of the finest Vice Presidents this country has ever had**], ”.[4]
3. “[**I've opposed the death penalty during all of my life**]. ”[5]
4. “[**I believe that we've got to get the power in the hands of the teachers, not the teachers' union**]”[6]

Examples 1 and 4 are claims which are asserting a certain policy, in example 2 an opinion is made and 3 depicts a claim in which the debater declares his stance against a policy.

3.2. Premise

In this annotation scheme we refer to the concept of data defined in Toulmin's model as premise. Premises are utterances asserted by debaters to justify a claim.

Similar to claims, we have also given some examples of indicators in the guidelines which help the annotators in finding premises in the debates. These indicators are: because, for example, for instance, so and etc. We have also mentioned in the guidelines that claims and premises sometimes come without any indicators, thus no part of these debates can be skipped for annotation.

Premises in political debates are also provided in several forms consisting facts, statistics, quotations, reports or examples, findings, physical evidence, or other reasoning methods. Several examples of the premises existent in the dataset are provided in the following.

5. [**Every estimate by this administration about the size of the deficit has been off by billions and billions of dollars**]. As a matter of fact, [**over 4 years, they've missed the mark by nearly $600 billion**]. [*We were told we would have a balanced budget in 1983*]. [*It was $200 billion deficit instead*]. And now we have a major question facing the American people as to whether we'll deal with this deficit and get it down for the sake of a healthy recovery. [*Virtually every economic analysis that I've heard of, including the distinguished Congressional Budget Office, which is respected by, I think, almost everyone, says that even with historically high levels of economic growth, we will suffer a $263 billion deficit*][7].
6. [*I have submitted an economic plan that I have worked out in concert with a number of fine economists in this country, all of whom approve it*], and [**believe that over a five year projection, this plan can permit the extra spending for needed refurbishing of our defensive posture**], that [**it can provide for a balanced budget by 1983 if not earlier**][8]

[2] Claims are written in **bold** and premises are written in *italics*. Component boundaries can be distinguished by [square brackets].

[3] Kennedy- 13 October 1960 in debate against Nixon

[4] Reagan- 21 October 1984 in debate against Mondale

[5] Dukakis- 13 October 1988 in debate against Bush

[6] Bush- 15 October 1992 in debate against Clinton and Perot

[7] Mondale, October 7, 1984 in debate against Reagan

[8] Reagan, October 28, 1980 in debate against Carter

7. [*The terrorism czar, who has worked for every president since Ronald Reagan, said, "Invading Iraq in response to 9/11 would be like Franklin Roosevelt invading Mexico in response to Pearl Harbor"*].[9]

   In examples 5 and 6 the debater tries to convince the audience of the credibility of his claim by giving as premise the confirmation of his claim by experts as references and in 7 the debater uses a quote as source to justify his claim.

   Premises can be stated as examples to justify the truth of a claim. In order to support a claim or come to a conclusion a debater may provide examples which will in our definition can be considered as premises. Examples of such premises are found in 8.

8. [**Race remains a significant challenge in our country**]. [**Unfortunately, race still determines too much**], [*often determines where people live*], [*determines what kind of education in their public schools they can get*], and, yes, [*it determines how they're treated in the criminal justice system*]. [*We've just seen those two tragic examples in both Tulsa and Charlotte*].[10]

In the examples above premises include facts such as the first three premises in example 8, or reports as instances which confirm the claim such as the last sentence in example 8. The warrant of the use of this premise is that it's plausible to generalize an occurrence of an event to deduct a rule.

4. Annotation Process

The annotation process is carried out by three annotators who have started annotating the argument components on the dataset using brat[11] annotation tool(Stenetorp et al., 2012) set up on an standalone server.

Training of the annotators is performed as a cyclic chain of test and trial. Firstly, a guideline was prepared to describe the annotation scheme with definitions and examples from the dataset, with common structures and controversial examples to train the annotators for the process. Figure 2 depicts this cyclic process.

Test periods were dedicated to the evaluation of each annotator based on their assigned annotation by two experts in the field of argumentation. After receiving the evaluation, their common mistakes were discussed and they were asked to revise their annotations. After each evaluation period the guidelines were revised by adding more examples for clarification according to the common mistakes made by annotators such as recognition of component boundaries, clarification of the differences between components. Trial periods are designed to observe the accordance of the annotation performed by annotators to the annotation by an expert and computing the inter-annotator agreement for the annotation.

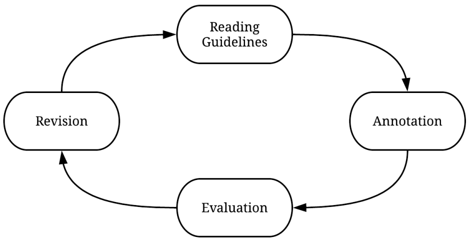

Figure 2: Cyclic process of annotating the dataset

5. Dataset Challenges

We confronted a number of challenges in the annotation process of the dataset.

The first challenge we encountered was the unstructured nature of the spoken language compared to written language such as persuasive essays which is used in (Stab and Gurevych, 2014). Lack of writing structure leads to lack of argument structure in this dataset. Thus, many claims exist in this dataset for which the candidates do not provide any justification. This is usually not the case in written argumentative essays.

Secondly, the task of annotating the arguments of a text is highly context-dependent, therefore annotators are supposed to decide tagging mainly based on the context. The problem in this dataset is that in a dialogue there are no specific boundaries to limit the context. In a persuasive essay, one might suggest that the annotators read a whole paragraph or even the whole essay to decide whether a sentence is a claim in the context of the essay as done in (Stab and Gurevych, 2014). This is however not possible in the debates dataset. In the guidelines we roughly suggested that the annotators read each speech turn first and then start annotating it.

In the guidelines we have provided some indicator phrases to facilitate the process of finding claims and premises in the text. Although these indicators can be helpful in most cases, there are cases in which they could mislead the annotators. Occurrence of these indicators in a sentence is not a guarantee that argument components are used in the sentence, this is shown in example 9.

9. IFILL: In that case, we'll move on to domestic matters. And this question, I believe, goes to Senator – to Vice President Cheney. The Census Bureau
   CHENEY: I think it goes to Senator Edwards.[12]

Another one of the main challenging characteristics of political debates is that there is no specific topic for each of the arguments that occur in the debate separately. Thus, unlike the work of (Stab and Gurevych, 2014), we cannot annotate any part of the document as a major claim for the current arguments of the debate and specify which claims are asserted under the topic of one major claim. We therefore have confined our annotation scheme for the annotation of components to claims and premises.

[9] Kerry, September 30, 2004 in debate against Bush

[10] Clinton, 26 September 2016 in debate against Trump

[11] http://brat.nlplab.org/

[12] October 5, 2004

6. Annotation results

In this study, the reliability of annotated dataset is measured by an inter-annotator agreement quantifier. This section provides some statistics and according to the partial annotation of the dataset until now[13].

We provide the distribution of annotated components roughly both at sentence level and token level.

	Sentence-Level%	Token-Level%
Claim	42.80	28.38
Premise	44.36	32.72
Not Annotated	13.49	38.93

Table 2: The percentage of annotated components at sentence and token level.

Table 2 illustrates the percentages of argument components annotated in the dataset. The sum percentage exceeds a hundred percent which indicates that there are less than 0.1% sentences which contain more than one component annotation. Since this difference is trivial we are going to present the IAA in measure of number of sentences containing argument components as an approximation.

Due to the large size of the corpus, instead of having each annotator annotate every transcript available in the corpus, we have decided that each transcript in the dataset should be annotated by two annotators.

Thus, the inter-annotator agreement is computed according to the agreement between each pair of the annotators and then reported as the average of these different reliability values.

In Table 3, we compute the agreement between boundaries of annotated vs non-annotated parts. The first column in table 3 shows the observed agreement value which is a quantity showing agreement over the annotation without chance correction. The second column uses a chance corrected inter-annotator reliability named Kripendorff's α introduced in (Krippendorff, 2004). The agreement of 0.6209 is the average agreement between different annotators on same transcripts.

Annotator Pairs	Average Observed Agreement	Agreement based on Kripendorff's α
A and B	0.8649	0.6791
A and B	0.8624	0.6463
A and C	0.7737	0.5374
Average	0.8336	0.6209

Table 3: Agreements between annotated-non annotated sentences.

The Confusion Probability Matrix (CPM) in Figure 3 illustrates the normalized disagreement between the annotators on sentence-level between different components.

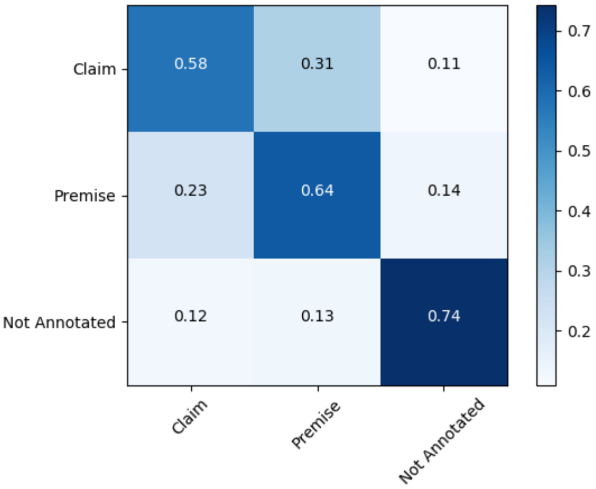

Figure 3: Normalized Confusion Probability Matrix of argument components

It can be inferred from the results in Figure 3 that the disagreement between premise and claim is very high. This disagreement often occurs due to the fact that some claims provided by debaters can be inferred as premises. For instance, in example 10, the sentence [**Communism is the enemy of all religions**] is provided to support the claim why [**we who do believe in God must join together**]. Although it is a claim which should be proven, there is no justification why that's a true statement and the debater uses it to support his previous claim therefore might be misannotated as a premise.

10. Nixon-Kennedy- October 13 1960:
    [**Communism is the enemy of all religions**]; and [**we who do believe in God must join together**]. [**We must not be divided on this issue**]

7. Related work

The political domain allows for intuitive applications of the argument mining framework with the final aim of detecting fallacies, persuasiveness degree and coherence in the candidate's argumentation. (Lippi and Torroni, 2016a) address the problem of argument extraction, and more precisely claim detection, over a corpus based on the 2015 UK political election debates. They aim to study the impact of the vocal features of speech on the claim detection task. The Internet Argument Corpus[14] (IAC) (Walker et al., 2012) collects the posts from `4forums.com`, a website for political debate. The debates have been annotated for argumentative markers like degrees of agreement with a previous post, cordiality, audience direction, combativeness, assertiveness, emotionality of argumentation, and sarcasm. (Duthie et al., 2016) apply AM methods to detect the presence and polarity of ethotic arguments from UK parliamentary debates.[15] The authors also investigate how

[13] The results in this section only covers 30 percent of the whole corpus which was annotated during the first cycle in the process.

[14] `http://nlds.soe.ucsc.edu/software`

[15] `http://arg.tech/Ethan3Train`, `http://arg.tech/Ethan3Test`

their results can be visualized to support user understanding.[16] (Naderi and Hirst, 2015) show how features based on embedding representations can improve discovering various frames in argumentative political speeches. They propose a corpus of speeches from the Canadian Parliament, and they examine the statements with respect to the position of the speaker towards the discussed topic (pro, con, or no stance). In (Menini et al., 2018), we address the relation prediction task on political speeches in monological form, where there is no direct interaction between the opponents. We created a corpus, based on the transcription of speeches and official declarations issued by Nixon and Kennedy during 1960 Presidential campaign, of argument pairs annotated with the support and attack relations.[17] None of these approaches considers the annotation of argument components (i.e., premises and claims) on a corpus of political debates, which is the object of our contribution.

8. Conclusion

In this paper, we have discussed the issue of identifying premises and claims in political debates. More precisely, we have presented the annotation guidelines we defined to train the annotators. The guidelines define what we mean by premises and claims in the context of political debates, and provide several examples to show instances of these two argument components. Since we are in the annotation phase of our dataset, we discussed the challenges we faced in training the annotators for this non trivial task, and we provide some statistics about the current status of our resources.

Future work includes the finalization of the annotation process of the dataset of political debates, and the definition of suitable NLP methods for the automatic identification of these argument components and the relations between them.

9. References

Duthie, R., Budzynska, K., and Reed, C. (2016). Mining ethos in political debate. In *COMMA*.

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, 38:787–800.

Lippi, M. and Torroni, P. (2016a). Argument mining from speech: Detecting claims in political debates. In *AAAI*.

Lippi, M. and Torroni, P. (2016b). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2018). Never retreat, never retract: Argumentation analysis for political speeches. In *AAAI-2018*.

Naderi, N. and Hirst, G. (2015). Argumentation mining in parliamentary discourse. In *CMNA*.

Stab, C. and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Toulmin, S. (2003). The uses of argument. 1958. *Cambridge: Cambridge UP*.

Walker, M., Tree, J. F., Anand, P., Abbott, R., and King, J. (2012). A corpus for research on deliberation and debate. In *LREC*.

Zarefsky, D. (1993). *Lincoln, Douglas, and slavery: In the crucible of public debate*. University of Chicago Press.

[16] https://goo.gl/P9fyzi

[17] https://dh.fbk.eu/resources/political-argumentation