Never retreat, Never retract: Argumentation Analysis for Political Speeches*

Stefano Menini,[1,3] Elena Cabrio,[2] Sara Tonelli,[1] Serena Villata[2]
[1]Fondazione Bruno Kessler, Trento, Italy
[2] Université Côte d'Azur, CNRS, Inria, I3S, France
[3]University of Trento, Italy
{menini, satonelli}@fbk.eu
{elena.cabrio, serena.villata}@unice.fr

Abstract

In this work, we apply argumentation mining techniques, in particular relation prediction, to study political speeches in monological form, where there is no direct interaction between opponents. We argue that this kind of technique can effectively support researchers in history, social and political sciences, which must deal with an increasing amount of data in digital form and need ways to automatically extract and analyse argumentation patterns. We test and discuss our approach based on the analysis of documents issued by R. Nixon and J. F. Kennedy during 1960 presidential campaign. We rely on a supervised classifier to predict argument relations (i.e., support and attack), obtaining an accuracy of 0.72 on a dataset of 1,462 argument pairs. The application of argument mining to such data allows not only to highlight the main points of agreement and disagreement between the candidates' arguments over the campaign issues such as Cuba, disarmament and health-care, but also an in-depth argumentative analysis of the respective viewpoints on these topics.

Introduction

In recent years, the analysis of argumentation using Natural Language Processing methods, so-called *argument mining* (Green et al. 2014), has gained a lot of attention in the Artificial Intelligence research community and has been applied to a number of domains, from student essays (Stab and Gurevych 2014) to scientific articles (Teufel, Siddharthan, and Batchelor 2009) and online user-generated content (Wachsmuth et al. 2014; Habernal and Gurevych 2015). However, while some of these approaches have been proposed to detect claims in political debates, e.g. (Lippi and Torroni 2016a; Naderi and Hirst 2015), little attention has been devoted to the prediction of relations between arguments, which could help historians, social and political scientists in the analysis of argumentative dynamics (e.g., supports, attacks) between parties and political opponents. For example, this analysis could support the study of past political speeches and of the repercussions of such claims over time. It could also be used to establish relations with the current way of debating in politics. In order to find argumentation patterns in political speeches, typically covering a wide range of issues from international politics to environmental challenges, the application of computational methods to assist scholars in their qualitative analysis is advisable.

In this work, we tackle the following research question: *To what extent can we apply argument mining models to support and ease the analysis and modeling of past political speeches?* This research question breaks down into the following subquestions:

- Given a transcription of speeches from different politicians on a certain topic, how can we automatically predict the relation holding between two arguments, even if they belong to different speeches?
- How can the output of the above-mentioned automated task be used to support history and political science scholars in the curation, analysis and editing of such corpora?

This issue is investigated by creating and analysing a new annotated corpus for this task, based on the transcription of discourses and official declarations issued by Richard Nixon and John F. Kennedy during the 1960 US Presidential campaign. Moreover, we develop a relation classification system with specific features able to *predict support* and *attack* relations between arguments (Lippi and Torroni 2016b), distinguishing them from unrelated ones. This argumentation mining pipeline ends with the visualization of the resulting graph of the debated topic using the OVA$^+$ tool.[1]

The main contributions of this article are *(1)* an annotated corpus consisting of 1,462 pairs of arguments in natural language (around 550,000 tokens) covering 5 topics, *(2)* a feature-rich Support Vector Machines (SVM) model for relation prediction, and *(3)* an end-to-end workflow to analyse arguments that, starting from one or more monological corpora in raw text, outputs the argumentation graph of user-defined topics.

The paper is organized as follows: first, we provide some basics about the argument mining pipeline and discuss the related work. Then, we describe the Presidential election campaign corpus, and we detail our experimental setting together with the obtained results. Finally, we present the visualization interface, and we analyse the argumentation patterns emerging from it. Conclusions end the paper.

*E. Cabrio and S. Villata have received funding from EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 690974 (MIREL project).

[1]http://ova.arg-tech.org/

Argument Mining

In the last years, the increasing amount of textual data published on the Web has highlighted the need to process it in order to identify, structure and summarize this huge amount of information. Online newspapers, blogs, online debate platforms and social networks, but also legal and technical documents provide a heterogeneous flow of information where natural language arguments can be identified and analyzed. The availability of such data, together with the advances in Natural Language Processing and Machine Learning, supported the rise of a new research area called *argument mining*. The main goal of argument mining is the automated extraction of natural language arguments and their relations from generic textual corpora, with the final goal to provide machine-readable structured data for computational models of argument and reasoning engines.

Two main stages have to be considered in the typical argument mining pipeline, from unstructured natural language documents towards structured (possibly machine-readable) data (Lippi and Torroni 2016b):

Argument extraction: The goal of the first stage of the pipeline is to detect arguments in natural language texts. Referring to standard *argument graphs* (Dung 1995), the retrieved arguments will thus represent the nodes in the final argument graph returned by the system. This step may be further split in two different stages such as the extraction of arguments and the further detection of their boundaries. Many approaches have recently tackled such challenge adopting different methodologies such as Support Vector Machines (Palau and Moens 2011), Naïve Bayes classifiers (Biran and Rambow 2011), Logistic Regression (Levy et al. 2014).

Relation prediction: The second stage of the pipeline consists in constructing the argument graph to be returned as output of the system. The goal is to predict what are the relations holding between the arguments identified in the first stage. This is an extremely complex task, as it involves high-level knowledge representation and reasoning issues. The relations between the arguments may be of heterogeneous nature, like attack, support or entailment (Cabrio and Villata 2013; Stab and Gurevych 2016). This stage consists also in predicting the internal relations among the argument's components, such as the connection between the premises and the claim (Biran and Rambow 2011; Stab and Gurevych 2014).

To tackle these challenging tasks, high-quality annotated corpora are needed, as those proposed in (Reed and Rowe 2004; Palau and Moens 2011; Levy et al. 2014; Stab and Gurevych 2014), to be used as gold standard data. For an exhaustive overview of argument mining techniques and applications, we refer the reader to (Peldszus and Stede 2013; Lippi and Torroni 2016b).

Few approaches apply part of the argument mining pipeline to political debates. Among them, (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; (Egan, Siddharthan, and Wyner 2016) propose an automatic approach to summarize political debates, starting from a political debates corpus (Walker et al. 2012); (Duthie, Budzynska, and Reed 2016) apply argument mining methods to mine ethos arguments from UK parliamentary debates, and (Naderi and Hirst 2015) show how features based on embedding representations can improve discovering various frames in argumentative political speeches. However, to the best of our knowledge, there are no approaches in the argument mining literature that tackle the problem of relation prediction over political speeches. The most important feature of such speeches is their monological nature, with unaligned arguments, while debates are typically characterised by two interlocutors answering each other. This leads to more implicit attack and support relations between the arguments put forward by the candidates.[2] Applying the argument mining pipeline, and more precisely, the relation prediction stage to such speeches is the goal of our contribution.

Corpus Extraction and Annotation

Since no data for this task are available, we collect the transcription of speeches and official declarations issued by Nixon and Kennedy during 1960 Presidential campaign from The American Presidency Project.[3] The corpus includes 881 documents, released under the NARA public domain license, and more than 1,6 million tokens (around 830,000 tokens for Nixon and 815,000 tokens for Kennedy). We select this document collection because of its relevance from a historical perspective: the 1960 electoral campaign has been widely studied by historians and political scientists, being the first campaign broadcast on television. The issues raised during the campaign shaped the political scenario of the next decades, for example the rising Cold War tensions between the United States and the Soviet Union or the relationship with Cuba.

Dataset creation

In order to include relevant topics in the dataset, we asked a history scholar to list a number of issues that were debated during 1960 campaign, around which argumentation pairs could emerge. With his help, we selected the following ones: *Cuba*, *disarmament*, *healthcare*, *minimum wage* and *unemployment* (henceforth *topics*). We then extracted pairs of candidate arguments as follows. For each topic, we manually define a set of keywords (e.g., *[medical care, health care]*) that lexically express the topic. Then, we extract from the corpus all sentences containing at least one of these keywords, plus the sentence before and after them to provide some context: each candidate argument consists then of a snippet of text containing three consecutive sentences and a date, corresponding to the day in which the original speech was given during the campaign.

[2] In argument mining, a support is a statement (source of the relation) that underpins another statement (target of the relation). It holds between a target and a source statement if the source statement is a justification or a reason for the target statement.

[3] The American Presidency Project (`http://www.presidency.ucsb.edu/1960_election.php`)

In the following step, we combine the extracted snippets into pairs using two different approaches. Indeed, we want to analyse two different types of argumentations: those *between candidates*, and those emerging from the speeches uttered by the *same candidate* over time. In the first case, for each topic, we sort all the candidate arguments in chronological order, and then create pairs by taking one or more snippets by a politician and the one(s) immediately preceding it by his opponent. These data are thus shaped as a sort of indirect dialogue, in which Nixon and Kennedy talk about the same topics in chronological order. However, the arguments of a speaker are not necessarily the direct answer to the arguments of the other one, making it challenging to label the relation holding between the two.

In the second case, we sort by topic all the candidate arguments in chronological order, as in the previous approach. However, each candidate argument is paired with what the same politician said on the same topic in the immediately preceding date. These data provide information about how the ideas of Nixon and Kennedy evolve during the electoral campaign, showing, if any, shifts in their opinions. We follow these two approaches also with the goal to obtain a possibly balanced dataset: we expect to have more attack relations holding between pairs of arguments from different candidates, while pairs of arguments from the same candidate should be coherent, mainly supporting each other.

Through this pairing process, we obtain 4,229 pairs for the *Cuba* topic, 2,508 pairs for *disarmament*, 3,945 pairs for *health-care*, 6,341 pairs for *minimum wage*, and 2,865 pairs for *unemployment*, for a total of 19,888 pairs.

Annotation

From the pool of automatically extracted pairs, we manually annotate a subset of 1,907 pairs randomly selected over the five topics. Annotators were asked to mark if between two given arguments there was a relation of *attack* (see Example 1 on minimum wage), a relation of *support* (see Example 2 on disarmament) or if there was *no relation* (arguments are neither supporting, nor attacking each other, tackling different issues of the same topic).

Example 1

Nixon*: And here you get the basic economic principles. If you raise the minimum wage, in my opinion - and all the experts confirm this that I have talked to in the Government - above $1.15, it would mean unemployment; unemployment, because there are many industries that could not pay more than $1.15 without cutting down their work force. $1.15 can be absorbed, and then at a later time we could move to $1.25 as the economy moves up.*

Kennedy*: The fact of the matter is that Mr. Nixon leads a party which has opposed progress for 25 years, and he is a representative of it. He leads a party which in 1935 voted 90 percent against a 25-cent minimum wage. He leads a party which voted 90 percent in 1960 against $1.25 an hour minimum wage.*

Example 2

Nixon*: I want to explain that in terms of examples today because it seems to me there has been a great lack of understanding in recent months, and, for that matter in recent years, as to why the United States has followed the line that it has diplomatically. People have often spoken to me and they have said, Why can't we be more flexible in our dealings on disarmament? Why can't we find a bold new program in this area which will make it possible for the Soviet Union to agree? The answer is that the reason the Soviet Union has not agreed is that they do not want apparently to disarm unless we give up the right to inspection.*

Nixon*: People say, Now, why is it we can't get some imaginative disarmament proposals, or suspension of nuclear test proposals? Aren't we being too rigid? And I can only say I have seen these proposals over the years, and the United States could not have been more tolerant. We have not only gone an extra mile - we have gone an extra 5 miles - on the tests, on disarmament, but on everything else, but every time we come to a blocking point, the blocking point is no inspection, no inspection.*

The annotation guidelines included few basic instructions: if the statements cover more than one topic, annotators were asked to focus only on the text segments dealing with the chosen topic. Annotation was carried out by strictly relying on the content of the statements, avoiding personal interpretation. Examples of *attack* are pairs where the candidates propose two different approaches to reach the same goal, where they express different considerations on the current situation with respect to a problem, or where they have a different attitude with respect to the work done in the past. For example, in order to increase minimum wage, Nixon proposed to set it to 1.10$ per hour, while Kennedy opposed this initiative, claiming that 1.35$ should be the minimum wage amount. In this example, the opponents have the same goal, i.e., increase minimum wage, but their statements are annotated as an attack because their initiatives are different, clearly expressing their disagreement.

After an initial training following the above guidelines, 3 annotators were asked to judge a common subset of 100 pairs to evaluate inter-annotator agreement. This was found to be 0.63 (Fleiss' Kappa), which as a rule of thumb is considered a substantial agreement (Landis and Koch 1977). After that, each annotator judged a different set of argument pairs, with a total of 1,907 judgements collected. In order to balance the data, we discarded part of the pairs annotated with *no relation* (randomly picked).

Overall, the final annotated corpus[4] is composed of 1,462 pairs: 378 pairs annotated with *attack*, 353 pairs annotated with *support*, and 731 pairs where these relations do not hold. An overview of the annotated corpus is presented in Table 1.

Experiments on Relation Prediction

To facilitate the construction of argument graphs and support the argumentative analysis of political speeches, we propose an approach to automatically label pairs of arguments according to the relation existing between them, namely *support* and *attack*.

[4]The dataset is available at `https://dh.fbk.eu/resources/political-argumentation`

Topic	Attack	Support	No Relation
Cuba	38	40	180
Disarmament	76	108	132
Medical care	75	72	142
Minimum wage	125	80	107
Unemployment	64	53	170

Table 1: Topic and class distribution in the annotated corpus

Given the strategy adopted to create the pairs, the paired arguments may happen to be also unrelated (50% of the pairs are labeled with *no relation*). Therefore, we first isolate the pairs connected through a relation, and then we classify them as *support* or *attack*. Each step is performed by a binary classifier using specific features, which we describe in the following subsection. In the paper, we present the results obtained with the feature set that achieved the best performance on 10-fold cross validation.

Experimental setting

The *first step* concerns the binary classification of related and unrelated pairs. In this step the pairs annotated with support and attack have been merged under the *related* label. We first pre-process all the pairs using the Stanford CoreNLP suite (Manning et al. 2014) for tokenization, lemmatization and part-of-speech tagging. Then, for each pair we define three sets of features, representing the lexical overlap between snippets, the position of the topic mention in the snippet, as a proxy for its relevance, and the similarity of snippets with other related / unrelated pairs.

Lexical overlap: the rationale behind this information is that two related arguments are supposed to be more lexically similar than unrelated ones. Therefore, we compute *i)* the number of nouns, verbs and adjectives shared by two snippets in a pair, normalized by their length, and *ii)* the normalized number of nouns, verbs and adjectives shared by the argument subtrees where the topic is mentioned.

Topic position: the rationale behind this information is that, if the same topic is central in both candidate arguments, then it is likely that these arguments are related. To measure this, we represent with a set of features how often the topic (expressed by a list of keywords, see previous section on dataset creation) appears at the beginning, in the central part or at the end of each candidate argument.

Similarity with other related / unrelated pairs: the intuition behind this set of features is that related pairs should be more similar to other related pairs than to unrelated ones. For each topic, its merged *related* and *unrelated* pairs are represented as two vectors using a bag-of-words model. Their semantic similarity with the individual pairs in the dataset is computed through cosine similarity and used as a feature.

For classification, we adopt a supervised machine learning approach training Support Vector Machines with radial kernel using LIBSVM (Chang and Lin 2011).

In the *second step* of the classification pipeline, we take in input the outcome of the first step and classify all the pairs of related arguments as support or attack. We rely on a set of surface, sentiment and semantic features inspired by Menini and Tonelli (2016) and Menini et al. (2017). We adopt the **Lexical overlap** set of features used also for the first step, to which we add the features described below. In general, we aim at representing more semantic information compared to the previous step, in which lexical features were already quite informative.

Negation: this set of features includes the normalized number of words under the scope of a negation in each argument, and the percentage of overlapping lemmas in the negated phrases of the two arguments.

Keyword embeddings: we use word2vec (Mikolov et al. 2013) to extract from each argument a vector representing the keywords of a topic. These vectors are extracted using the continuous bag-of-word algorithm, a windows size of 8 and a vector dimensionality of 50.

Argument entailment: these features indicate if the first argument entails the second one, and vice-versa. To detect the presence of entailment we use the Excitement Open Platform (Magnini et al. 2014).

Argument sentiment: a set of features based on the sentiment analysis module of the Stanford CoreNLP suite (Socher et al. 2013) are used to represent the sentiment of each argument, calculated as the average sentiment score of the sentences composing it.

Additional features for lexical overlap, entailment and sentiment are obtained also considering only the subtrees containing a topic keyword instead of the full arguments. The feature vectors are then used to train a SVM with radial kernel with LIBSVM, like in the first classification step.

Evaluation

We test the performance of the classification pipeline using the 1,462 manually annotated pairs with 10-fold cross-validation. The first classification step separates the argument pairs linked by either an *attack* or a *support* relation from the argument pairs with *no relation* (that will be subsequently discarded). The purpose of this first step is to pass the related pairs to the *second step*. Thus, we aim at the highest precision, in order to minimise the number of errors propagated to the second step. Table 2 shows the results of the classification for the first step. We choose a configuration that, despite a low recall (0.23), scores a precision of 0.88 on the *attack/support* pairs, providing for the second step a total of 194 argument pairs.

	Unrelated	Attack/Support	Average
Precision	0.56	0.88	0.72
Recall	0.97	0.23	0.60
F1	0.71	0.36	0.65

Table 2: Step 1: classification of related / unrelated pairs

The second step classifies the related pairs assigning an *attack* or a *support* label. We provide two evaluations: we report the classifier performance only on the gold *attack* and *support* pairs (Table 3), and on the pairs classified as related in the first step (Table 4). In this way, we evaluate the classifier also in a real setting, to assess the performance of the

end-to-end pipeline.

	Attack	Support	Average
Precision	0.89	0.75	0.82
Recall	0.79	0.86	0.83
F1	0.84	0.80	0.82

Table 3: Step 2: classification of *Attack* and *Support* using only gold data.

	Attack	Support	Average
Precision	0.76	0.67	0.72
Recall	0.79	0.86	0.83
F1	0.77	0.75	0.77

Table 4: Step 2: classification of *Attack* and *Support* using the output of Step 1.

As expected, accuracy using only gold data is 0.82 (against a random baseline of 0.70), while it drops to 0.72 (against a random baseline of 0.51) in the real-world setting.

We also test a 3-class classifier, with the same set of features used in the two classification steps, obtaining a precision of 0.57. This shows that *support/attack* and *no relation* are better represented by using different sets of features, therefore we opt for two binary classifiers in cascade.

Notice that a comparison of our results with existing approaches to predict argument relations, namely the approach of (Stab and Gurevych 2016) on persuasive essays, cannot be fairly addressed due to huge differences in the complexity of the used corpus. With their better configuration, (Stab and Gurevych 2016) obtain an F1 of 0.75 on persuasive essays (that are a very specific kind of texts, human upperbound: macro F1 score of 0.854), and of 0.72 on microtexts (Peldszus and Stede 2013). The difference in the task complexity is highlighted also in the inter-annotator agreement. Differently from persuasive essays, where students are requested to put forward arguments in favour and against their viewpoint, in political speeches, candidates often respond to opponents in subtle or implicit ways, avoiding a clear identification of opposing viewpoints.

Error analysis

If we analyse the classifier output at topic level, we observe that overall the performance is consistent across all topics, with the exception of *minimum wage*. In this latter case, the classifier performs much better, with an accuracy of 0.94 in the second step. This is probably due to the fact that Kennedy's and Nixon's statements about minimum wage are very different and the discussion revolves around very concrete items (e.g., the amounts of the minimum wage, the categories that should benefit from it). In other cases, for example disarmament or Cuba, the speakers' wording is very similar and tends to deal with abstract concepts such as freedom, war, peace.

Furthermore, we observe that the classifier yields a better performance with argument pairs by the same person rather than those uttered by different speakers: in the first case, accuracy is 0.86, while in the second one it is 0.79 (Step 2).

Looking at misclassified pairs, we notice very challenging cases, where the presence of linguistic devices like rhetorical questions and repeated negations cannot be correctly captured by our features. Example 3 reports on a pair wrongly classified as *Support* belonging to the *health care* topic:

Example 3

Nixon: *Now, some people might say, Mr. Nixon, won't it be easier just to have the Federal Government take this thing over rather than to have a Federal-State program? Won't it be easier not to bother with private health insurance programs? Yes; it would be a lot simpler, but, my friends, you would destroy the standard of medical care.*

Kennedy: *I don't believe that the American people are going to give their endorsement to the leadership which believes that medical care for our older citizens, financed under social security, is extreme, and I quote Mr. Nixon accurately.*

Visualization and Analysis of the Argumentation Graphs

In this section, we describe how the results of our relation predition system are then used to construct the argumentation graphs about the debated topics.

Several tools have been proposed to visualize (and then reason upon) argumentation frameworks in the computational argumentation field, e.g., Carneades[5], GRAFIX[6], and ConArg2[7]. However, two main problems arise when trying to use such tools for our purposes: first, they are not tailored to long, natural language snippets (the usual names of arguments in computational argumentation are of the form arg_1), and second, they do not consider the possibility to identify specific argumentation schemes over the provided text. For all these reasons, we decided to rely upon a well-know tool called OVA$^+$ (Janier, Lawrence, and Reed 2014), an on-line interface for the manual analysis of natural language arguments. OVA$^+$ grounds its visualization on the Argument Interchange Format (AIF) (Chesñevar et al. 2006), allowing for the representation of arguments and the possibility to exchange, share and reuse the resulting argument maps. OVA$^+$ handles texts of any type and any length.

The last step of our argument mining pipeline takes in input the labeled pairs returned by the relation prediction module and translates this output to comply with the AIF format. This translation is performed through a script converting the `CSV` input file into `json` file to be load on OVA$^+$ through its online interface.[8] In this mapping, each argument is extracted in order to create an information node (I-node) (Chesñevar et al. 2006), and then, it is possible to create the associated locution node (L-node) and to specify the

[5] `http://carneades.github.io/`

[6] `https://www.irit.fr/grafix`

[7] `http://www.dmi.unipg.it/conarg/`

[8] The script and the argumentation graphs about the five topics in our corpus (both gold standard and system's output) are available at `https://dh.fbk.eu/resources/political-argumentation`

name of the speaker. The locution appears, preceded by the name of the participant assigned to it, and edges link the L-node to the I-node via an "Asserting" YA-node, i.e., the illocutionary forces of locutions, as in the Inference Anchoring Theory (IAT) model (Budzynska and Reed 2011). Supports or attacks between arguments are represented as follows, always relying upon the standard AIF model. A RA-node (*relation of inference*) should connect two I-nodes. To elicit an attack between two arguments, RA-nodes are changed into CA-nodes, namely *schemes of conflict*. Nodes representing the support and the attack relations are the "Default Inference" and the "Default Conflict" nodes, respectively. Figure 1 shows (a portion of) the argumentation graph resulting from the relation prediction step about the topic *minimum wage*, where three I-nodes (i.e., arguments) are involved in one support and one attack relation. The *Asserting* nodes connect each argument with its own source (e.g., K for Kennedy and N for Nixon).

OVA$^+$ allows users to load an analysis, and to visualize it. Given the loaded argumentation graph, the user is supported in analyzing the graph by identifying argumentation schemes (Walton, Reed, and Macagno 2008), and adding further illocutionary forces and relations between the arguments. This final step substantially eases the analysis process by historians and social scientists. Moreover, at the end of the analysis, OVA$^+$ permits to save the final argumentation graph on the user's machine (image or json file).

This graph-based visualization is employed to support political scientists and historians in analysing and modeling political speeches. This proves the usefulness of applying the argumentation mining pipeline over such kind of data: it allows users to automatically identify, among the huge amount of assertions put forward by the candidates in their speeches, the main points on which the candidates disagree (mainly corresponding to the solutions they propose to carry out or their own viewpoints on the previous administrations' effectiveness) or agree (mainly, general-purpose assertions about the country's values to promote).

In the following, we analyze the argumentative structure and content of two of the graphs resulting from the discussed topics (i.e., *minimun wage* and *health care*), highlighting main conflicting arguments among candidates, and other argumentative patterns. Note that this analysis is carried out on the proposed dataset, that contains a subset of all the speeches of the candidates, but gives a clear idea of the kind of analysis that could be performed by scholars on the entirety of the speeches. In general (and this is valid for all the analyzed graphs), we notice that the candidates almost always disagree either on the premises (e.g., who caused the problem to be faced) or on the proposed solutions (the minor claims).

Minimum wage. A widely discussed topic by both candidates was *minimum wage*, i.e., the bill to set the lowest remuneration that employers may legally pay to workers. It is worth noticing that the argumentation graph for the minimum wage corpus is rather complicated, and it highlights some main controversial issues. The candidates do not agree about the causes of the low minimum wage in 1960 in the US. More precisely, Kennedy attacks the fact that the administration supported an increase in the minimum wage by attacking Nixon's argument *"The misstatement: In the second debate Senator Kennedy said: The Republicans in recent years, not only in the last 25 years, but in the last 8 years, have opposed minimum wage. The facts : [. . .] The administration supported an increase in the minimum wage in 1955, and in 1957 urged legislation to extend minimum wage coverage to some 3 million additional workers, an extension which the Democratic-led Congress failed to approve. In 1960, this administration sought to extend minimum wage coverage to 3.1 million additional workers and indicated support of an increase in the minimum wage to $1.15 per hour."*. This argument is attacked from different perspectives, leading to a disagreement on the actions the administration carried out in the past years to deal with the minimum wage problem. For instance, as shown in Figure 1, Kennedy states that *"In the midthirties, 90 percent of the Republican Party voted against a 25-cent minimum wage. This summer, as your Congressman can tell you, in the House of Representatives , 90 percent of the Republicans voted against a minimum wage of $1.25 an hour, $50 a week for a 40-hour week, for a business that makes more than a million dollars a year, and Mr. Nixon called it extreme. He is frozen in the ice of his own indifference if I ever saw a Republican candidate who was"*. While we may say that this source of disagreement is about the causes of the minimum wage issue, another main source of disagreement is represented by the solutions proposed by the two candidates, which mainly differ regarding the amount of increase of the minimum wage and the coverage of the two respective bills. All these issues become evident and identifiable with ease in the resulting argumentation graph about the minimum wage topic.

Medical care. The problem of medical care for the elderly was a main problem in 1960, and this topic was widely discussed in the campaign. The resulting argumentation graph highlights some relevant argumentative patterns that are worth analyzing. In general, in the argumentation graphs we are analyzing, the support relation holds between arguments proposed by the same candidate, ensuring in this way a certain degree of coherence in their own argumentation. Interestingly, in the argumentation graph on the topic *medical care*, we can observe that a support relation holds between an argument from Kennedy and one from Nixon, i.e., *"Those forced to rely on surplus food packages should receive a more balanced, nourishing diet. And to meet the pressing problem confronting men past working age, and their families, we must put through an effective program of medical care for the aged under the social security system. The present medical care program will not send one penny to needy persons without further action by the Congress and the State legislatures."* supports *"N: We stand for programs which will provide for increased and better medical care for our citizens, and particularly for those who need it, who are in the older age brackets - and I will discuss that more a*

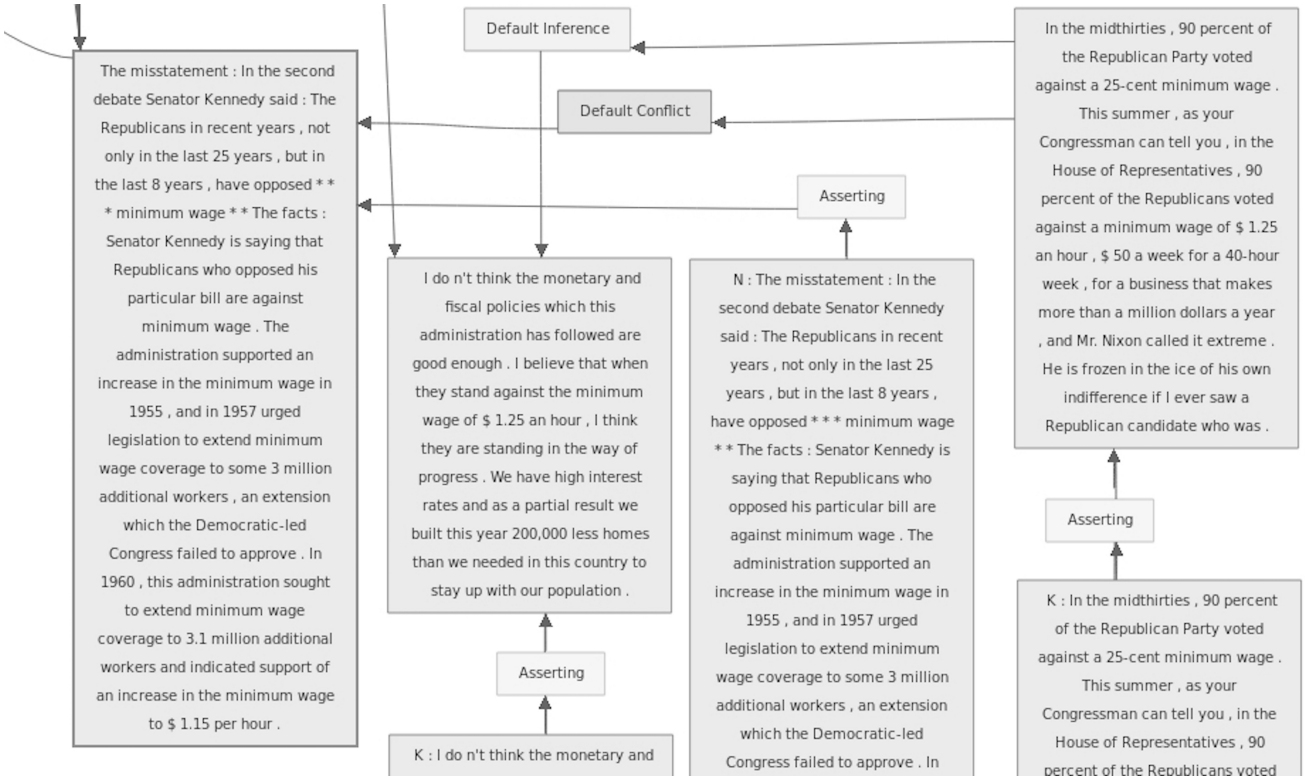

Figure 1: The argumentation graph about the topic *minimum wage* visualized through the OVA$^+$ tool.

little later. We stand for progress in all of these fields, and certainly, as I stand here before you, I am proud to be a part of that platform and of that program". These instances of support among candidates mostly concern general issues, i.e., a program of medical care for the elderly is needed.

In summary, our system allows to ease the detection of such argumentation patterns (i.e., topics on which both candidates agree, topics on which they disagree, topics on which they provide contradictory assertions) as well as to study how they connect with the other statements asserted in the speeches.

Conclusions

In this paper, an argumentation mining system for relation prediction has been presented and evaluated over a corpus of political speeches from the Nixon-Kennedy U.S. election campaign of 1960. The main advantage of the proposed approach is threefold. First of all, to the best of our knowledge, this is the first approach in argument mining targeting the relation prediction task in monological speeches, where interlocutors do not directly answer to each other. Our approach enables scholars to put together - and more importantly to connect - assertions from the two candidates across the whole political campaign. The output is thus an argumentation graph (one for each topic touched upon in the speeches) summarizing the candidates' own viewpoint and the respective position. Such graphs are intended to support researchers in history, social and political sciences, which must deal with an increasing amount of data in digital form and need ways to automatically extract and analyse argumentation patterns. Second, despite the complexity of the task that constituted a challenge in the annotation phase (and that can be observed in the reported examples), the results we obtained for relation prediction are in line with state-of-the-art systems in argument mining (Stab and Gurevych 2016). A third contribution of our work is a resource of 1,462 pairs of natural language arguments annotated with the relations of *support*, *attack*, and *no relation*. In the dataset, each argument is connected to the source and the date in which the speech was given.

As for future work, we face two major challenges. First, to improve the system performances, we need a finer-grained argument boundary definition. Namely, the goal is to identify within an argument its *evidences* and *claims*, so that the relations of support and attack may also be addressed towards these precise argument components. This would also have an impact on facilitating the work of scholars in the manual analysis of the argumentation graphs generated by our system. Second, we plan to evaluate the system with scholars in history and political sciences, who will be asked to judge not only the classification output but also the way in which it is displayed. We are currently working at a more interactive interface to display graphs with their textual content, so that users can select and visualize subgraphs according to the selected argumentative pattern.

References

Biran, O., and Rambow, O. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *Int. J. Semantic Computing* 5(4):363–381.

Budzynska, K., and Reed, C. 2011. Whence inference. Technical report, University of Dundee.

Cabrio, E., and Villata, S. 2013. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation* 4(3):209–230.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27.

Chesñevar, C. I.; McGinnis, J.; Modgil, S.; Rahwan, I.; Reed, C.; Simari, G. R.; South, M.; Vreeswijk, G.; and Willmott, S. 2006. Towards an argument interchange format. *Knowledge Eng. Review* 21(4):293–316.

Dung, P. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77(2):321–358.

Duthie, R.; Budzynska, K.; and Reed, C. 2016. Mining ethos in political debate. In *Proceedings of COMMA 2016*, 299–310.

Egan, C.; Siddharthan, A.; and Wyner, A. Z. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining*.

Green, N.; Ashley, K.; Litman, D.; Reed, C.; and Walker, V., eds. 2014. *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics.

Habernal, I., and Gurevych, I. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of EMNLP*, 2127–2137.

Janier, M.; Lawrence, J.; and Reed, C. 2014. OVA+: an argument analysis interface. In Parsons, S.; Oren, N.; Reed, C.; and Cerutti, F., eds., *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, 463–464. IOS Press.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174.

Levy, R.; Bilu, Y.; Hershcovich, D.; Aharoni, E.; and Slonim, N. 2014. Context dependent claim detection. In *Proceedings of COLING*, 1489–1500.

Lippi, M., and Torroni, P. 2016a. Argument mining from speech: Detecting claims in political debates. In *Proceedings AAAI*, 2979–2985.

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

Magnini, B.; Zanoli, R.; Dagan, I.; Eichler, K.; Neumann, G.; Noh, T.-G.; Pado, S.; Stern, A.; and Levy, O. 2014. The excitement open platform for textual inferences. In *Proceedings of ACL (System Demonstrations)*, 43–48.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL (System Demonstrations)*, 55–60.

Menini, S., and Tonelli, S. 2016. Agreement and disagreement: Comparison of points of view in the political domain. In *Proceedings of COLING*.

Menini, S.; Nanni, F.; Ponzetto, S. P.; and Tonelli, S. 2017. Topic-based agreement and disagreement in us electoral manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2928–2934. Stroudsburg, PA: Association for Computational Linguistics.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Naderi, N., and Hirst, G. 2015. Argumentation mining in parliamentary discourse. In *Proceedings of the 15th Workshop on Computational Models of Natural Argument (CMNA-2015)*.

Palau, R. M., and Moens, M. 2011. Argumentation mining. *Artif. Intell. Law* 19(1):1–22.

Peldszus, A., and Stede, M. 2013. From argument diagrams to argumentation mining in texts: A survey. *IJCINI* 7(1):1–31.

Reed, C., and Rowe, G. 2004. Araucaria: Software for argument analysis, diagramming and representation. *Int. Journal on Artificial Intelligence Tools* 13(4):961–980.

Socher, R.; Perelygin, A.; Wu, J. Y.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, volume 1631, 1642.

Stab, C., and Gurevych, I. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of EMNLP*, 46–56.

Stab, C., and Gurevych, I. 2016. Parsing argumentation structures in persuasive essays. *CoRR* abs/1604.07370.

Teufel, S.; Siddharthan, A.; and Batchelor, C. R. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP*, 1493–1502.

Wachsmuth, H.; Trenkmann, M.; Stein, B.; and Engels, G. 2014. Modeling review argumentation for robust sentiment analysis. In *Proceedings of COLING*, 553–564.

Walker, M.; Tree, J. F.; Anand, P.; Abbott, R.; and King, J. 2012. A corpus for research on deliberation and debate. In *Proceedings of LREC*, 812–817. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1643.

Walton, D.; Reed, C.; and Macagno, F. 2008. *Argumentation Schemes*. Cambridge University Press.