ARG-tech
Centre for Argument Technology

ENS DE LYON

June 6 2016 - August 19 2016

Automated intertextual correspondence mining
Social media reacting to political argumentation

Maria Boritchev

Supervised by
Chris Reed

Abstract

This study presents an approach to automated mining of intertextual correspondences between political argumentation and social media reactions. In other words, the aim is to identify allusions, quotations, calque, parody, etc. in the particular context of political discourse argumentation and social media reactions regarding it. During public events (such as elections), it is important for politicians to know what is the audience's reaction to the argument strategies they deploy. Identifying the most efficient (reaction-rising) techniques can not only give the politicians the possibility to adjust their discourse but also give an attentive audience clues to critical thinking. Addressing the intertextual correspondence problem requires the creation of two annotated argumentative discourse corpora and the design of a mining algorithm specific to this framework. This work develops the processes that were used.

Contents

Introduction

Analysing political argumentation and its effects on the targeted audience is a complex problem linking approaches specific to argument mining and general natural language processing ones. Addressing this problem is a matter of major interest in argument studies, as it extends the existing socio-cultural knowledge of political discourse impact and permits a more mindful approach to major political events such as 2016 U.S presidential elections. Since 2008, Chris Reed and his colleagues from ARG-tech (Center for Argument Technology[1]) of the University of Dundee have been developing argumentation analysis tools and models. Among those, Inference Anchoring Theory (IAT) provides a framework for studying argumentation in dialogues, and the OVA+ interface[2] permits a visualization of textual arguments analysis. Using OVA+ to achieve a full IAT analysis of 2016 U.S. primaries campaign debates and corresponding Reddit comments threads, we create corpora needed to state the question of intertextual correspondence mining. Then, combining several Natural Language Processing approaches, we design an algorithm to address it.

This document is my first year of Master's degree in Computer Science at the École Normale Supérieure de Lyon internship report. This internship was done at ARG-tech and supervised by Professor Chris Reed. The aim of this work is to mine intertextual correspondences between Reddit users' comments and U.S. candidates argumentation during primaries debates, using and extending argumentation analysis tools and models provided by ARG-tech.

This report first presents the question in its scientific and socio-linguistic context. Then it focuses on the development of the corpora we need to conduct our research and the specific model choices that are made. Finally, we present an algorithmic design for addressing the intertextual correspondence problem.

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

1 Motivation of the problem

U.S. presidential elections stand among the major expected events in international politics of 2016. As President Obama is finishing his second term and hence cannot be re-elected, these elections oppose candidates from major political parties (Democratic and Republican) that have to be chosen during primary elections. To present new candidates to the voters and influence their decisions, highly publicized debates are organized. These debates give the candidates the possibility to show their best sides to the public while arguing on main points of their future policy as well as on more personal details.
Primaries debates are organized between members of the same parties, by mass media outlets. Candidates and moderators (journalists, TV correspondents) meet before an audience of citizens. Debates follow strict rules: they are more of a strictly guided interview than of a free dialogue. The main moderator asks questions, introduces follow-ups and guides the discussion. Other moderators can join in the questioning. Time-limits can vary from one debate to another; each candidate has one minute or 90 seconds to answer questions and 30 or 45 seconds for follow-ups and rebuttals – if their name has been mentioned in another's candidate answer.

Reddit[3] is a social media and news aggregation, web content rating and discussion English-speaking website. Reddit's users can submit content such as text posts or direct links. They can also comment on other people's submissions. The submissions are organized by areas of interest called "subreddits", including topics such as news, science, politics, gaming, movies, fitness, food, among many others. Inside subreddits, which are organized in a forum fashion, users can access "threads", one for each subject. For example, entering the subreddit "/r/politics"[4] gives access, among other diverse threads, to the "October 13 DNC Primary Debate - During-debate Discussion Megathread"[5]. Redditors can then post comments inside the thread, and respond back and forth in a conversation-tree of comments (see Figure 1).

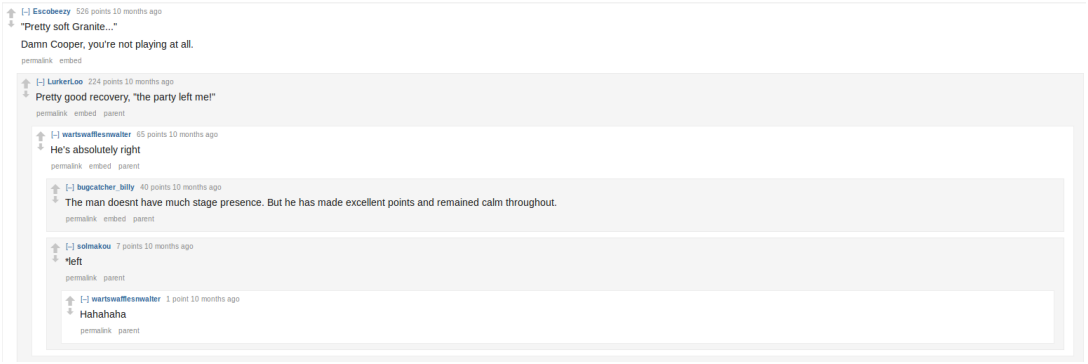

Figure 1: An example of a Reddit conversation-tree of comments.

As to Reddit users themselves, according to Reddit's Audience and Demographics page[6], 54% are from the Unites States, therefore potential voters in the upcoming elections. Yet, research results published by the Pew Internet Project [Duggan and Smith, 2013] show that only 6% of adults active on the internet are redditors, the most represented population among those being young (age 18 to 29) males. Reddit is a social media less widely spread in the population

[3] https://www.reddit.com/

[4] https://www.reddit.com/r/politics/

[5] https://www.reddit.com/r/politics/comments/3onts4/october_13_dnc_primary_debate_duringdebate/

[6] https://reddit.zendesk.com/hc/en-us/articles/205183225-Audience-and-Demographics

than Facebook or Twitter. Thus, using Reddit as part of the framework for our study gives us the possibility to refine our results, as redditors come from sociologically close backgrounds.

The aim of this study is to provide some quantification and qualification of the social embedding of political argumentation in this particular framework. How do candidates convince voters? How to measure the effect of candidates' argumentation on the worldwide audience? How do future voters react to presidential debates? More specifically, *what links can be found between presidential primaries debates and social reactions regarding them? Are reactions specific to one candidate? To one topic? To one moment of the debate?*
We do not have the ambition here to show the whole picture or answer these questions extensively, but to develop techniques for intertextual correspondence mining starting from this particular case study. As Inference Anchoring Theory [Budzynska and Reed, 2011] provides a framework for connecting dialogical structures with argumentative structures and both primaries presidential debates and Reddit conversation-trees of comments present obvious dialogical structures, we can use IAT to extract argumentative structures from these linked sources. Then, we are able to automatically mine the intertextual correspondences between these structures, giving a starting point to social embedding of political argumentation investigation.

2 Related work

Argument mining is a relatively new challenge in the field of corpus-based discourse analysis. It concerns automated identification of argumentative structures that can be found within the discourse. For each argument we find premises, conclusion and argumentative scheme (as *modus ponens*, argument from authority, etc.) but also the relationships between pairs of arguments in the text (argument/subargument, argument/counterargument, etc.).
Applications of argument mining go from general improvement of information retrieval and extraction to helping visualization and summarization of arguments for users. Techniques and results can be incorporated in domains such as computer-supported peer review or computerized essay grading. Argument mining is a problem combining approaches from fields such as natural language processing (NLP), theories of semantics, pragmatics, discourse and argumentation, artificial intelligence.
As argument mining inherits NLP approaches, it requires the creation and annotation of high-quality corpora of argumentative discourse. Using "real-life" texts ensures that the research does not only characterize its authors specific speech and argumentation characteristics. Argument mining corpora are composed of argumentative texts, at least partly annotated by hand. The corpora on which this study is based have been built and annotated during the internship (see the corpora-building rules bellow).
Mining intertextual correspondences is a problem related to automatic detection and attribution of quotations [Pareti et al., 2013], paraphrases [Brockett and Dolan, 2005] and even textual entailment recognition [Bentivogli et al., 2009]. All the challenges cited here have been studied using machine learning and various natural language processing techniques. However combining them on specific data is yet another task.

Previously, ARG-tech's work has been based on mediation corpora, then on moral debates. Each time, fully-annotated corpora have been developed. They can be found, among many others, in the Argument Web database (AIFdb)[7]. Building the presidential debates and Reddit comments corpora and integrating them to AIFdb has required learning and mastering the Inference Anchoring Theory, a major model and tool developed by ARG-tech.

[7] `http://corpora.aifdb.org/`

3 Methodology

Inference Anchoring Theory developed by ARG-tech is a model and tool for linking dialogical and argumentative structures. It allows the expression of relations between logic and argumentation theory. It is specifically tailored to handle argumentative discourse, i.e. discourse giving reasons in support of the claims it contains, in order to influence an audience. Before giving the formalization of IAT, let us consider some introducing examples.

Example 1. A simple dialogue:
Alice: Q.
Bob: *Why Q?*
Alice: P.

In this example, the dialogue presents a logical structure: a conclusion Q, a premise P, and the application of an implicit rule $P \rightarrow Q$. As we do not have the lexical content of the dialogue, we can't specify the rule; it could be purely logical, as *modus ponens*, or argumentative (in which case, it will be called an *argumentation scheme*) as *argument from consequence* [Walton et al., 2008]. A classical logic equivalent of this dialogue could be the following inference tree:

$$\frac{P}{Q}\, P \rightarrow Q$$

But as we are working with dialogue and therefore with dialogical structures [Walton, 1984], using this representation induces a huge loss of information. Indeed: Alice's utterance "Q" gives the right to Bob to ask "Why Q?" because Alice has committed herself to "Q". Formulated in an IAT fashion, Bob follows the dialogical rule [Mcburney and Parsons, 2002] stipulating that *challenging* a person is allowed after this person has made an *assertion.*

Example 2. A small Reddit comment tree:
Redditor1: *Every American should be a capitalist.*
Redditor2: *Why?*
Redditor1: *Our country was built on capitalism.*

Example 2 shows a "real-life" occurrence of our simple dialogue from Example 1. Figure 2 displays the IAT diagram corresponding to this dialogue. It is drawn using OVA+ following the process described below.

Analysing the excerpt and drawing the diagram requires first to find argumentative *speech acts*: scraps of text implementing the argumentation. Then, we need to identify the relation between the *propositional content* of the speech acts. This relation can be `inference` (when one argument follows from the other), `conflict` or `rephrase` (when one argument is a restating of the other).
Once the propositional contents are linked, we can link the corresponding speech acts following the chronological order: on the argumentative side, conclusion can come before the premise, thus making the inference go backwards, but on the speech act side, the transition goes the other way, echoing the text's structure.
Then we link argument and dialogical structures by determining the intended *illocutionary forces*: `Asserting`, `Challenging`, `Questioning`, `Arguing`, `Disagreeing`, etc.

IAT diagrams follow (as a general rule) a 3 columns structure (see Figure 2). On the right hand side, there are five nodes representing applications of dialogical rules: three speech acts

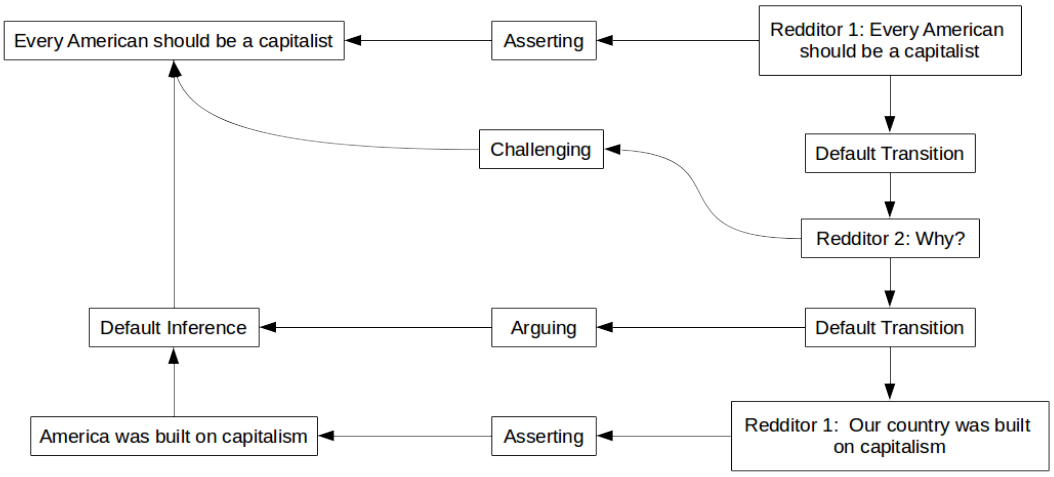

Figure 2: IAT: Linking argument and dialogical structures.

(`Redditor1: Every American should be a capitalist`) and two dialogical transitions – `Default Transition`. On the left hand side, there are nodes representing argument structures: two propositional contents of speech acts, which can be either identical to the original speech acts or reformulated to be self-sufficient for understanding (`America was built on capitalism`) and one associated relation (`Default Inference` here, but could be `Default Conflict` or `Default Rephrase`). In the middle, the nodes represent illocutionary forces, linking argument and dialogical structures (here, `Asserting`, `Challenging`, `Arguing`).

OVA+ is a drag-and-drop interface designed to realise IAT analysis. All IAT diagrams found in this report have been drawn using OVA+, but drawing assistance is not the only feature of this tool. OVA+ drawn diagrams can be exported in a `JSON` (JavaScript Object Notation) format (see Annex for full `JSON` of the example above). Moreover, it gives the possibility to export *segments* only: the argument scraps resulting from the full IAT analysis. For example, the `JSON` segmentation corresponding to Figure 2 is the following:

```
{"nodes": [
    {
      "text": "every American should be a capitalist",
      "type": "I"
    },
    {
      "text": "America was built on capitalism",
      "type": "I"
    }
}
```

`JSON segmentation`

The intertextual correspondences mining problem addressed here can now be stated in a more proper way. Our goal is to mine intertextual correspondences (allusion, quotation, calque, pastiche, parody, etc.) by mapping full IAT analysis given segments from Reddit comments threads to the ones extracted from the 2016 U.S. primaries campaign debates.

4 Mining the arguments

As said above, an argument mining approach requires the creation and annotation of argumentative discourse corpora. For the study presented in this report, we needed to create two corpora, one for the Primaries Presidential Debates and the other for Reddit Comments. To obtain a homogeneous result, IAT annotator's guidelines are decided before the annotation starts. These guidelines can only be written after extensively studying the textual data.

4.1 Presidential debates corpus

The American Presidency Project (APP[8]) is a web archive gathering American presidency related documents. It is the leading source on the internet in this domain, now containing 118,793 documents. Among other various presidency-related resources, APP offers transcripts of all presidential debates from 1960 to 2016, both from primary and general elections. For 2016's elections, we for now have access to transcripts from 21 primary elections debates, 9 Democratic and 12 Republican ones (only main debates being considered). These debates transcripts constitute the starting point for the Presidential Debates corpus.
An issue is that a debate lasts at least one hour and a half. As our aim is to mine arguments that would produce the most explicit Reddit reaction, we do not want to analyse the whole of all debates. Our hypothesis is *Panem et circenses*: bread and games; the audience is more likely to react when there's action on stage. In the context of political debates, action on stage corresponds to high dialogical activity moments.
There can be several ways to define high dialogical activity (HDA) moments. In IAT terms, HDA moments of the debate's transcription can be identified by looking at the diagram: links between arguments of different participants get more complex as edges between nodes cannot be untangled anymore (see Figure 3).

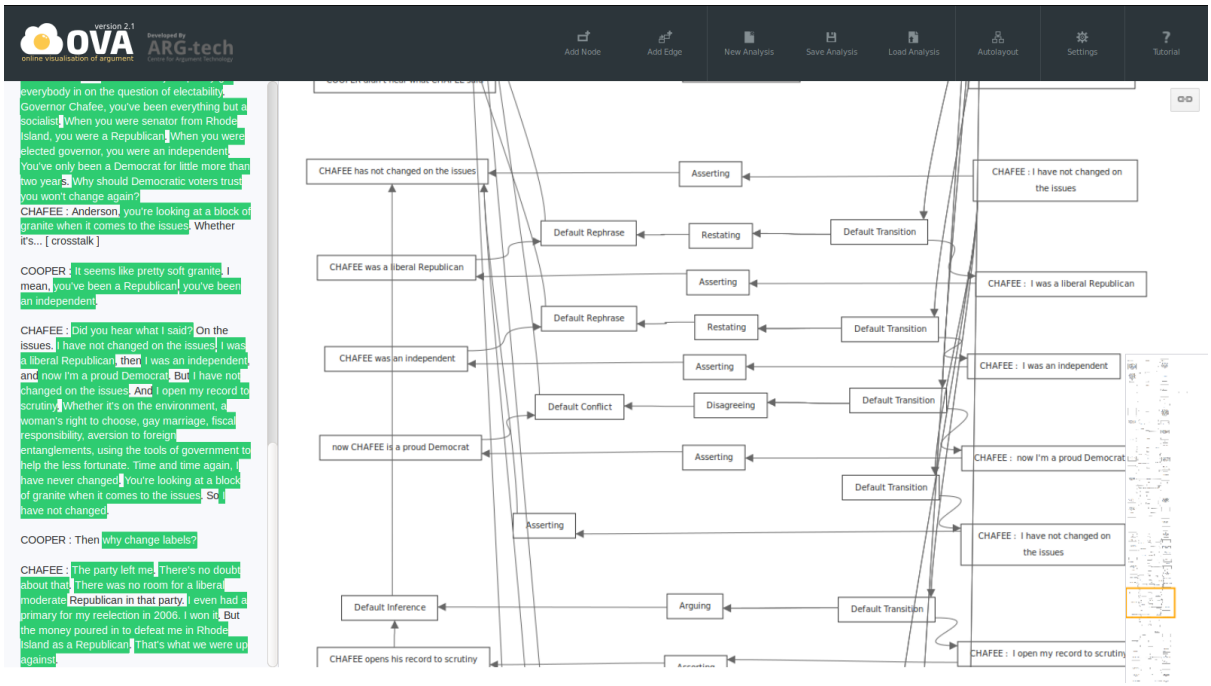

Figure 3: IAT: High dialogical activity analysis excerpt.

[8] `http://www.presidency.ucsb.edu/index.php`

After working on the transcripts and the videos of the debates, the following method of HDA excerpts mining was developed.

Definition 1 (High dialogical activity center)**.** *A* high dialogical activity center *is a keyword encountered in the transcription of a dialogue such that its presence indicates a HDA moment.*

For example, in Figure 3, we can see that `crosstalk` is a HDA center. Similarly, we can determine that encountering `agree` and `disagree` often precedes a HDA moment.
One slightly more subtle type of HDA center can be found while considering the debates' process. As explained before, the usual procedure goes like this: a moderator asks a question, a participant (a candidate) answers, a moderator asks another question, and so on. But when a participant mentions the name of another participant during his turn, the mentioned participant has the right to respond and can sometimes (when the debate is animated) take this right without waiting for an invitation from the moderator to do so. This situation can be formulated the following way: when two participants' turns follow one another without a moderator's turn, the first participant's name (indicating the beginning of the turn) is a HDA center.
Once we have an idea of what are HDA centers, we need to define the HDA parameter: HDA moments' half span.

Definition 2 (High dialogical activity parameter)**.** *The* high dialogical activity parameter k *is the number of speech turns prior and posterior to the HDA center that should be considered to select the HDA moment corresponding to the HDA center.*

Example 3. Taking the text used in the IAT analysis presented in Figure 3 (see Appendix B) and setting k to 1, we get the following excerpt:

> *COOPER : We're going to have a lot more on these issues. But I do want to just quickly get everybody in on the question of electability. Governor Chafee, you've been everything but a socialist. When you were senator from Rhode Island, you were a Republican. When you were elected governor, you were an independent. You've only been a Democrat for little more than two years. Why should Democratic voters trust you won't change again?*
>
> *CHAFEE : Anderson, you're looking at a block of granite when it comes to the issues. Whether it's... [crosstalk]*
>
> *COOPER : It seems like pretty soft granite. I mean, you've been a Republican, you've been an independent.*

Once the HDA parameter and the HDA centers are defined, two questions naturally appear. How many different sorts of HDA centers should we consider? What should the k be?
Making the value of k and the sorts of HDA centers vary (see Annex for quantification), we came to the conclusion that the four previously defined kinds of HDA centers were necessary but sufficient for our study. As to the HDA parameter, $k = 3$ gives excerpts long enough not not lose context necessary for credible – not over-interpretative – argument analysis.

Quite naturally, it may happen that two HDA excerpts mined using this method overlap. In this case, we merge the overlapping excerpts into a new bigger one. From this decision follows the upper bound on k: as we increase the value of k, excerpts get longer, thus the chances that two excerpts overlap grow, which leads to even longer final excerpts. And here come the problems: as handy as OVA+ is, annotators can only handle diagrams of reasonable size. Therefore, measuring the risk of loosing some information, we stick to $k = 3$.

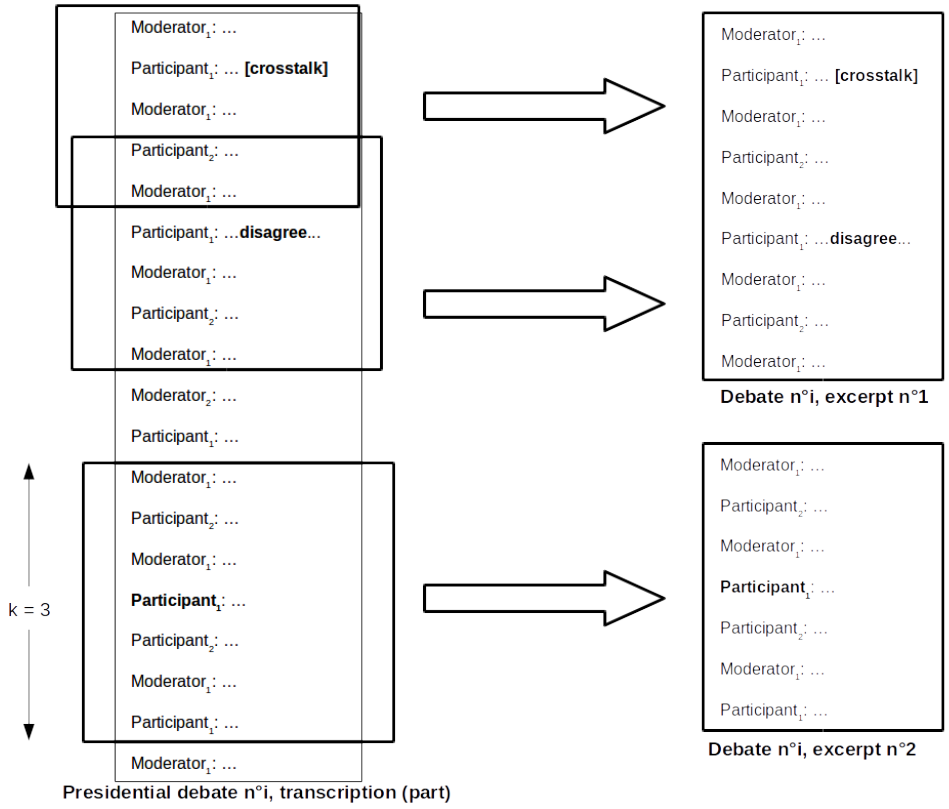

Figure 4: HDA moments extraction.

After mining the HDA moments using this method (see Figure 4), we obtain the Presidential Debates Corpus, subdivided by debate and then by excerpts. For example, the first democratic debate of the October 13 2015 has been subdivided in 13 HDA moments (see Appendix A for the first excerpt).
For correspondence mining purposes, each excerpt has been timestamped: knowing the beginning time of the debate, we have the beginning and end time of each excerpt.

4.2 Reddit comments corpus

Once the Presidential Debates Corpus is built, we build the Reddit related comments one. First, we identify all the threads corresponding to the primaries debates. The next step is the process bellow:

1. For each debate, select the corresponding thread.
2. Sort the comments by time-stamp (oldest on top).
3. Remove all comments having no children.
4. Remove all comments trees beginning with irony or wordplay (rethoric structures are not handled by IAT).
5. Keep comments trees classified by excerpts (time-stamp identification), discard all others.

The time-stamp classification of each comment is done by comparing the comment's time-stamp with the beginning and end (+1 minute) times of each excerpt. The additional minute is an account for Redditors reaction time: it is impossible to instantly react by written comment. The selection rules above may seem to eliminate an overwhelmingly large proportion of the data. In fact, as the threads we are working with are composed of more than 10,000 comments (15,754 comments for the "October 13 DNC Primary Debate - During-debate Discussion Megathread"), this selection process is not so data-expensive.

5 Architecture

The argument mining process we conducted in the field of social embedding of political argumentation study has led to the creation of two segments corpora. Figure 5 gives the scheme of this work.

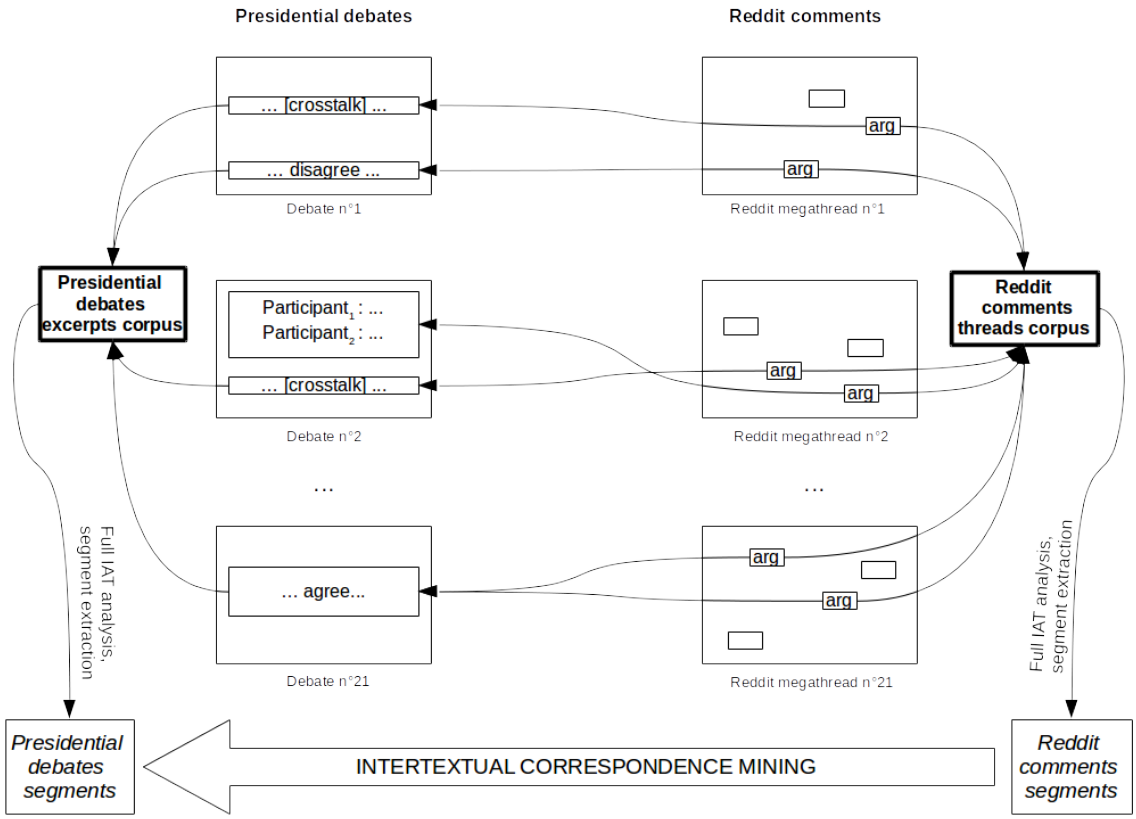

Figure 5: The big picture.

Starting from the full transcripts of all primaries presidential debates, we narrow the scope of our analysis to high dialogical activity excerpts which we mine around high dialogical activity centers – words said during the debate, words used to describe action happening on stage, indicators of speech turns. These excerpts are assembled in a text corpus, timestamped, then undergo full IAT analysis. JSON segmentation then gives us the presidential debates segments. After the presidential debates data is processed, Reddit comments are filtered to leave only the ones that not only can be analysed with IAT but also correspond (according to time-stamping)

to the previously mined presidential debates excerpts. The remaining comments are assembled in a text corpus and undergo full IAT analysis. JSON segmentation then gives us the Reddit comments segments.
Having the two segment corpora, we need to come up with a method for intertextual correspondence mining of Reddit comments regarding presidential debates.

6 Algorithm design

The intertextual correspondence mining problem can now be stated as following:

Input: $R = \{r_0, \ldots, r_i, \ldots, r_n\}$ array of n Reddit comments segments, $D = \{d_0, \ldots, d_j, \ldots, d_m\}$ array of m presidential debates segments.

Output: $C{=}(c_{i,j})$, matrix of the correspondence coefficients between r_i and d_j.

The first idea for computing $(c_{i,j})$ (given i and j) is to have a look at the content of r_i and d_j. As the data we are considering here is not composed of random words but has been selected in a specific way, we can use lexical similarity: the names of the presidential candidates appear in both segments.

Definition 3 (Speakers set)**.** *We call* speakers set *of a segment t ($t \in R$ or D) $S(t) = \{s_0, \ldots, s_l\}$ composed of the speakers names recognized in t.*

Example 4. *For the segment "Hillary Clinton knows Bernie Sanders' gun control record isn't his strong suit" ($\in R$), the speakers set is $\{CLINTON, SANDERS\}$. For the segment "CLINTON thinks what Senator Sanders is saying certainly makes sense in the terms of the inequality that we have" ($\in D$), the speakers set is the same.*

Once the speakers sets of all segments are computed, we can define the speakers similarity.

Definition 4 (Speakers similarity)**.** *We call* speakers similarity *of two segments r_i and d_j ($r_i \in R$ and $d_j \in R$) the function Sp defined as follows:*

$$Sp(r_i, d_j) = \begin{cases} 0 & \text{if } S(r_i) \cup S(d_i) = \varnothing \\ \dfrac{|S(r_i) \cap S(d_i)|}{|S(r_i) \cup S(d_i)|} & \text{otherwise, } |A| \text{ being the number of elements of the set } A. \end{cases}$$

We observe, by definition, that $Sp(t, t) = 1$ – a segment has full speaker similarity with itself.

After defining the speaker similarity feature, specific to our problem, we can consider related NLP questions. Among those, we focus on *Word Similarity*. Two approaches are usually employed and combined to address this problem: corpus-based methods and ontology-based ones. As we are working with texts, words have to be considered within their context – for example ambiguous words can have different meanings depending on the sentence they are used in. Getting back to our correspondence problem, we would like to compute word similarity between the "most important" words of r_i and d_j. To identify these words, we consider the *term frequency–inverse document frequency* (tf-idf).
Computing the term frequency of a word in a text permits to quantify how often it is used and therefore, how important it is. Unfortunately, some very frequent words (such as "a", "the", "he", "she", etc.) are also identified by this method, making it inefficient. This is why inverse document frequency is considered: as the word is in a text which is part of a corpus, we can compute the inverse of the frequency of the word in the corpus. Multiplying the two statistics, we obtain the information we needed to identify the most important and generally rare words in each segment.

Definition 5 (Term frequency)**.** *Let w be a word of a (non-empty) segment t in a corpus C.* Term frequency *of w in t is defined as*

$$Tf(w,t) = \frac{|\{v \in t, v = w\}|}{|\{v \in t\}|}.$$

Definition 6 (Inverse document frequency)**.** *Let w be a word in a segment t in a corpus C.* Inverse document frequency *of w in C is defined as*

$$Idf(w,C) = \begin{cases} 0 & \text{if } w \text{ never appears in any segment of } C \\ \log \dfrac{|s, s \in C|}{|s \in C, w \in s|} & \text{otherwise.} \end{cases}$$

Inverse document frequency is logarithmically scaled to fit the term frequency order of magnitude – a corpus may be composed of many one-word texts.

Definition 7 (Term frequency–inverse document frequency)**.** *Let w be a word in a segment t in a corpus C.* tf-idf *of w in t is defined as*

$$\textit{tf-idf}(w,t,C) = Tf(w,t) \cdot Idf(w,C).$$

Computing the sum of $\textit{tf-idf}(w, d_j, D)$ for each w in r_i, we obtain a coefficient reflecting how important are words of r_i in d_j with respect to the corpus D. Interpreting it regarding our data, this coefficient is high when Reddit comments contain direct quotations extracted from the presidential debates.

Back to the Word Similarity problem, let us consider ontology-based approaches. By definition, an ontology is a set of concepts and categories in a subject area or domain that shows their properties and the relations between them. For NLP, one of the most famous ontologies is WordNet[9], a lexical database for the English language created and maintained by the Cognitive Science Laboratory of Princeton University. All words in WordNet have the word `entity` as common ancestor, and are linked by hyponym/hypernym relations. `entity` has three *hyponyms* ("children" of more specific meaning): `physical entity#1` – "physical entity" in its first sense, `abstraction#6` – "abstraction" in its sixth sense, `thing#8`. However, WordNet is not organised in a tree fashion, as some words can have several *hypernyms* ("ancestors" of wider meaning). Figure 6 illustrates a fragment of WordNet's hierarchy. Several measures for semantic similarity between words and concepts are implemented in WordNet; here, we choose to focus on path based ones.

Indeed, computing the length of the shortest path linking a concept (a sense of a word) c_1 to a concept c_2 in WordNet hierarchy gives an indication on how semantically close the two concepts are.

Example 5. *The length of the shortest path linking* `compartment#2` *to* `instrumentality#3` *(see Figure 6) is* 5*.*

As our data is not semantically annotated, we need to decide on a way to solve the word sense ambiguity problem before using WordNet given semantic similarities. Word Sense Disambiguation is a NLP question on its own, therefore we decided not to address it in this study, and get around the problem while keeping it in mind.

[9] `https://wordnet.princeton.edu/`

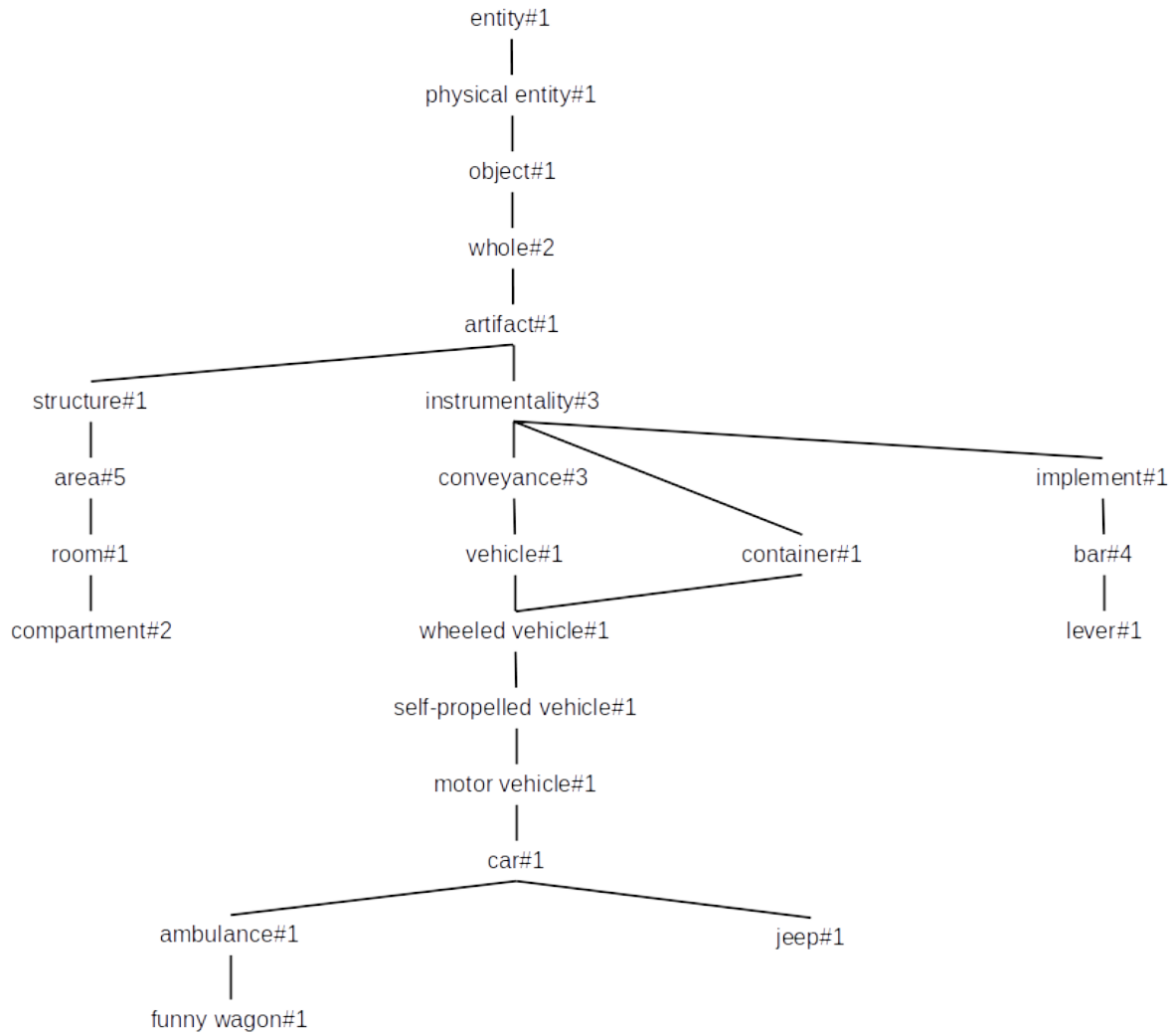

Figure 6: WordNet hierarchy (part.).

Definition 8 (Semantic similarity)**.** *We define the* semantic similarity *of two words* w_1 *and* w_2 *using the following Semsim function:*

$$Semsim(w_1, w_2) = 1 - \frac{\min_{i,j}\{length(path(w_1\#i, w_2\#j))\}}{\max_{v,k}\{length(path(\texttt{entity}, v\#k))\}}$$

i being in range of w_1's senses, j being in range of w_2 ones, v being a word in WordNet (not `entity`*), i being in range of v's senses.*

Computing the sum of $Semsim(u, v)$ for each pair u in r_i, v in d_j, we obtain a coefficient reflecting how semantically close words of r_i are regarding words in d_j. If $u = v$, $Semsim(u, v) = 1$, so if $r_i = d_j$, this coefficient is equal to the squared length of the segment. As our goal is to obtain comparable correspondence coefficients $c_{i,j}$, we need to normalize the semantic similarity coefficient by dividing it by $length(r_i) \cdot length(d_j)$. Applying this idea to all the considered coefficients, computation of the $c_{i,j}$ looks as following:

$$c_{i,j} = w_1 \cdot Sp(r_i, d_j) + \frac{w_2}{length(r_i)} \cdot \sum_{u \in r_i} tf\text{-}idf(u, d_j, D) + \frac{w_3}{length(r_i) \cdot length(d_j)} \cdot \sum_{u \in r_i, v \in d_j} Semsim(u, v)$$

r_i being in R, d_j in D. w_1, w_2 and w_3 are weights, to be experimentally adjusted, reflecting the mutual importance of each term of the sum.

The task left is experimental: setting the weights and the corresponding threshold – the minimum value of $c_{i,j}$ above which r_i is considered to be in intertextual correspondence with d_j. As the implementation is ongoing, we hope that our work will give us a new way of studying the social embedding of political argumentation.

Conclusion

Constructing the two corpora has been an important part of this project. Once the process was settled, addressing the intertextual correspondence mining problem became a task anchored in the concrete examples we had. Using external tools and integrating them to the framework of our study has permitted to make the first steps towards a resolution of some social embedding of political argumentation asked questions. The solution presented in this report could now be enlarged either by adding other audience samples or considering other countries elections, but also by integrating previous political argumentation related studies. *A Functional Analysis of 2012 US Presidential Primary Debates* ([Glantz et al., 2013]) focuses on the argumentation of presidential primary debates candidates as they use it against one another. Integrating this point of view in the correspondence mining process may enrich the approach, and the list of future enhancements that is given here is clearly not close to an end. Correspondence mining is and stays an open research question.

Acknowledgements

I would first like to thank my internship advisor Professor Chris Reed for directing my research and giving me the possibility to discover a field of work which was completely new to me. I would also like to thank Katarzyna Budzynska for introducing me to team research work and for the insight that she brought to the project.
I am really grateful to the whole ARG-tech team – Rory Duthie, Martín Pereira Fariña, Mathilde Janier, Barbara Konat, Marcin Koszowy, John Lawrence, Elaine McIntyre, Alison Pease, Mark Snaith, Jacky Visser – for the various help and advice they offered me. Working with them was rewarding, pleasant and enlightening.

References

[Bentivogli et al., 2009] Bentivogli, L., Clark, P., Dagan, I., and Giampiccolo, D. (2009). The sixth pascal recognizing textual entailment challenge. *Proceedings of TAC*, 9:14–24.

[Brockett and Dolan, 2005] Brockett, C. and Dolan, W. B. (2005). Support vector machines for paraphrase identification and corpus construction. In *Proceedings of the 3rd International Workshop on Paraphrasing*, pages 1–8.

[Budzynska and Reed, 2011] Budzynska, K. and Reed, C. (2011). How dialogues create arguments. In Van Eemeren, F. e. a., editor, *Proceedings of the 7th Conference of the International Society for the Study of Argumentation (ISSA 2010)*.

[Duggan and Smith, 2013] Duggan, M. and Smith, A. (2013). 6% of online adults are reddit users. *www.pewresearch.org*.

[Glantz et al., 2013] Glantz, M., Benoit, W. L., and Airne, D. (2013). A functional analysis of 2012 us presidential primary debates. *Argumentation and Advocacy*, 49:275.

[Mcburney and Parsons, 2002] Mcburney, P. and Parsons, S. (2002). Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of Logic, Language, and Information*.

[Pareti et al., 2013] Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R., and Koprinska, I. (2013). Automatically detecting and attributing indirect quotations. In *EMNLP*, pages 989–999.

[Walton, 1984] Walton, D. (1984). *Logical Dialogue-Games and Fallacies*. University Press of America.

[Walton et al., 2008] Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.

Appendices

A Transcript excerpt

COOPER: You don't consider yourself a capitalist, though?
SANDERS: Do I consider myself part of the casino capitalist process by which so few have so much and so many have so little by which Wall Street's greed and recklessness wrecked this economy? No, I don't. I believe in a society where all people do well. Not just a handful of billionaires. [applause]
COOPER: Just let me just be clear. Is there anybody else on the stage who is not a capitalist?
CLINTON: Well, let me just follow-up on that, Anderson, because when I think about capitalism, I think about all the small businesses that were started because we have the opportunity and the freedom in our country for people to do that and to make a good living for themselves and their families. And I don't think we should confuse what we have to do every so often in America, which is save capitalism from itself. And I think what Senator Sanders is saying certainly makes sense in the terms of the inequality that we have. But we are not Denmark. I love Denmark. We are the United States of America. And it's our job to rein in the excesses of capitalism so that it tt run amok and doesn't cause the kind of inequities we're seeing in our economic system. But we would be making a grave mistake to turn our backs on what built the greatest middle class in the history...
COOPER: Senator Sanders?
CLINTON: ...of the world. [applause]
SANDERS: I think everybody is in agreement that we are a great entrepreneurial nation. We have got to encourage that. Of course, we have to support small and medium-sized businesses. But you can have all of the growth that you want and it doesn't mean anything if all of the new income and wealth is going to the top 1 percent. So what we need to do is support small and medium-sized businesses , the backbone of our economy, but we have to make sure that every family in this country gets a fair shake...
COOPER: We're going to get...
SANDERS: ...not just for billionaires.
COOPER: We're going to have a lot more on these issues. But I do want to just quickly get everybody in on the question of electability. Governor Chafee, you've been everything but a socialist. When you were senator from Rhode Island, you were a Republican. When you were elected governor, you were an independent. You've only been a Democrat for little more than two years. Why should Democratic voters trust you won't change again?
CHAFEE: Anderson, you're looking at a block of granite when it comes to the issues. Whether it's... [crosstalk]
COOPER: It seems like pretty soft granite. I mean, you've been a Republican, you've been an independent.
CHAFEE: Did you hear what I said? On the issues. I have not changed on the issues. I was a liberal Republican, then I was an independent, and now I'm a proud Democrat. But I have not changed on the issues. And I open my record to scrutiny. Whether it's on the environment, a woman's right to choose, gay marriage, fiscal responsibility, aversion to foreign entanglements, using the tools of government to help the less fortunate. Time and time again, I have never changed. You're looking at a block of granite when it comes to the issues. So I have not changed.
COOPER: Then why change labels?
CHAFEE: The party left me. There's no doubt about that. There was no room for a liberal moderate Republican in that party. I even had a primary for my reelection in 2006. I won it. But the money poured in to defeat me in Rhode Island as a Republican. That's what we were up against.

B Corresponding IAT diagram

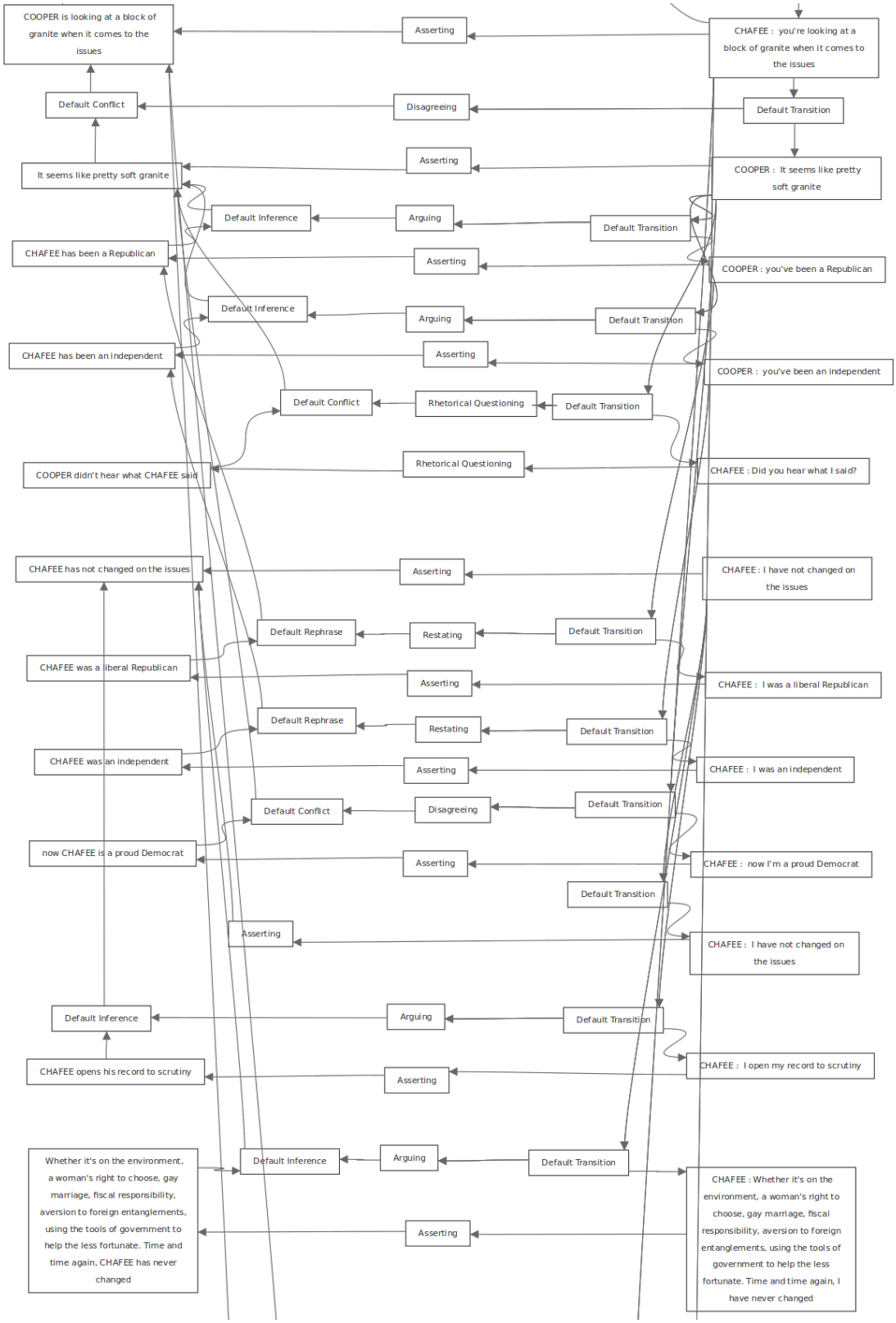

Figure 7: IAT diagram (part.).