1

Processing Natural Language Argumentation

Katarzyna Budzynska, Serena Villata

ABSTRACT. Although natural language argumentation has attracted the attention of philosophers and rhetoricians since Greek antiquity, it is only very recently that the methods and techniques of computational linguistics and machine learning have become sufficiently mature to tackle this extremely challenging topic. *Argument mining*, the new and rapidly growing area of natural language processing and computational models of argument, aims at automatic recognition of argument structures in large resources of natural language texts. The goal of this chapter is to familiarise the reader focused on formal aspects of argumentation with this approach, and to show how argument structures, e.g. those studied in abstract argumentation frameworks, can be extracted, providing a bridge between mathematical models and natural language. To this end, we first describe the typical argument mining pipeline and related tasks, and then present in more detail a specific example of work in this area.

1 Introduction

Argument mining[1] (see e.g. [Moens, 2013; Peldszus and Stede, 2013; Lippi and Torroni, 2015a] for an overview) aims to develop methods and techniques which allow for automation of the process of identification and extraction of argument data from large resources of natural language texts. This task is split into two parts: linguistic and computational. The linguistic part consists of the analysis (called annotation) of natural language texts of a selected type of discourse (called a genre) and the association of tags with segments of the texts (elementary discourse units, EDUs). For example, if we want to automatically recognise the categories of premise and conclusion, then the annotator analysing the text should assign the tags "premise" and "conclusion" to identified EDUs. The set of texts with tags, i.e. the corpus, constitutes an input for the computational part of the process. The tagging is used to teach algorithms to automatically identify categories or aspects of argumentation such as "premise" and "conclusion". The techniques are based on either a structural or a statistical approach. In the structural approach, a linguist analyses part of the corpus and identifies patterns between the tagged categories and various cues present, e.g. at the linguistic surface of the text. For instance, the linguist may observe that whenever the word "since" occurs, the following segment of

[1] Also referred to or associated with argumentation mining, computational argumentation or debating technologies.

text is a premise. The resulting set of such patterns is called a grammar. In the statistical approach, the task of recognising patterns is automatised, i.e. a machine learns these patterns using statistical correlations. The developed or trained algorithms can then be used to mine arguments in large resources of natural language texts such as Wikipedia.

The area of argument mining continues and extends two other areas of natural language processing – sentiment analysis and opinion mining[2], which have proved to be very successful and important both academically and commercially. *Sentiment analysis* develops techniques for fast, automatic processing of texts, such as posts on social media, in order to detect positive and negative sentiments expressed on the Internet towards various products, companies, or people. These techniques are widely used in predicting stock market trends. *Opinion mining* moves from searching for general attitudes ("I like Mercedes") to recognising specific beliefs and opinions ("Mercedes cars are reliable"). Its techniques have found an important application in media analysis, where automation significantly accelerates analysis of natural language texts as compared to manual analysis and consequently enables more accurate summaries of people's opinions, e.g. what they think about a candidate for the presidency. Argument mining expands these research tasks further, allowing us to extract valuable information not only about what attitudes and opinions people hold, but also about the arguments they give in favour of (supporting arguments) and against (conflicting arguments) these attitudes and opinions.

The chapter consists of two parts. First, after a brief introduction to the history (Section 1.1) and background (Section 1.2) of argument mining and the types of arguments that are mined (Section 1.3), in Section 2 we present the typical pipeline of Natural Language Processing (NLP) methods and techniques employed for automatic recognition and extraction of argumentation. In Section 3, we then describe in detail an example of work in argument mining [Cabrio and Villata, 2012b; Cabrio and Villata, 2013] in order to provide a better understanding of how this pipeline is actually used to process argumentation in natural language texts. This chapter does not aim to present the area in a detailed and exhaustive way – we recommend interested readers to have a look at already existing, excellent surveys such as [Moens, 2013; Peldszus and Stede, 2013; Lippi and Torroni, 2015a]. Our goal here is to provide a roadmap of argument mining which shows the most typical approaches, makes links to the work presented in other chapters of this handbook, and offers pointers to a variety of papers for further relevant reading.

[2] In the literature, the terms "sentiment analysis" and "opinion mining" either are treated as names referring to two different areas of study or are used interchangeably. Here we follow the first convention, in order to highlight the nuanced distinction between people's general attitude (sentiment) and specific attitude (opinion) to products, institutions, persons, etc.

1.1 A short overview of studies on natural language argumentation

The communicative phenomenon of natural language argumentation has attracted attention since the very beginning of science. Aristotle popularised argument studies in his *Rhetoric* [Aristotle, 2004], distinguishing three means of persuasion: *logos*, i.e. the argumentation or propositional content of the message; *ethos*, i.e. the speaker's character or credibility; and *pathos*, the emotions of the audience. Since that time rhetoric has continued to expand and evolve over many centuries.

At the end of the nineteenth century, the success of mathematical logic drew attention away from natural reasoning to formal reasoning. The dominance of this approach, however, did not last long – in the 1950s it was criticised by a number of researchers, such as Perelman & Olbrechts-Tyteca [1958] and Toulmin [1958], who pointed out the limitations of deduction in modelling real-life communication. Since this turning point, the theory of natural language argumentation has become an important area of research in the disciplines of philosophy, linguistics, discourse analysis, communicative studies, and rhetoric, focusing on topics such as argument structure and schemes, fallacies, enthymemes, dialogue games, counter-argumentation of rebuttals, and undercutters (see e.g. [Groarke, 1996; van Eemeren *et al.*, 2014] for a detailed overview). To facilitate and foster the exchange of ideas amongst members of this interdisciplinary community, two international associations were founded: the International Society for the Study of Argumentation (ISSA), which has organised conferences every four years since 1986[3]; and the Ontario Society for the Study of Argumentation (OSSA), which has held conferences every two years since 1995[4]. Since 2015, another large biennial conference, the European Conference on Argumentation (ECA), has been organised[5]. These conferences gather hundreds of academics from all around the world and various disciplines.

Natural language argumentation has recently begun to attract increasing attention from researchers with a computationally oriented perspective, in particular from the communities of Computational Models of Argument (COMMA)[6] and Workshops on Argument Mining (co-organised with NLP conferences such as the Association for Computational Linguistics, ACL)[7]. The year 2014 was an important milestone for argument mining, resulting in many publications and events, e.g. (1) "From Real Data to Argument Mining" at the 12th ArgDiaP workshop, Polish Academy of Sciences, Warsaw, Poland, 23-24 May 2014 (22 papers); (2) the First Workshop on Argumentation Mining at the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA, June 26, 2014 (18 papers); (3) the SICSA Workshop on Argument Mining: Perspectives from Information Extraction, Information Retrieval

[3] http://cf.hum.uva.nl/issa/
[4] http://scholar.uwindsor.ca/ossaarchive/
[5] http://ecargument.org/
[6] http://www.comma-conf.org/
[7] http://argmining2016.arg.tech/

and Computational Linguistics, Centre for Argument Technology ARG-tech, Dundee, Scotland, 9-10 July 2014 (about 25 participants); (4) "Frontiers and Connections between Argumentation Theory and Natural Language Processing", Bertinoro (Forli-Cesena), Italy, 21-25 July 2014 (about 30 participants); (5) the Second Workshop on Argumentation Mining, 53rd Annual Meeting of the Association for Computational Linguistics, Denver, Colorado, USA, June 04, 2015 (16 papers); (6) "Debating Technologies", Dagstuhl Seminar, December 14-18, 2015 (about 25 participants); (7) "Arguments in Natural Language: The Long Way to Analyze the Reasons to Believe and the Reasons to Act" at the First European Conference on Argumentation, Lisbon, 9-12 July 2015; (8) "Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments", Dagstuhl Seminar, April 17-22, 2016 (about 40 participants); and (9) the Third Workshop on Argumentation Mining at the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, August 2016 (20 papers and about 70 participants). In 2016, two tutorials on argument mining were accepted for presentation at the top conferences in artificial intelligence and computational linguistics: the 25th International Joint Conference on Artificial Intelligence (IJCAI-16), New York, NY, USA, 11 July 2016[8] and the annual meeting of the Association for Computational Linguistics (ACL 2016), Berlin 7 August 2016[9]. In 2017 three further tutorials have been delivered at the main graduate school dedicated to language and communication: the 29th European Summer School in Logic, Language, and Information (ESSLLI 2017), Toulouse (France), 17-28 July, 2017. Finally, projects on argument mining have been recognised as important by several funding institutions, e.g. the project "Robust Argumentation Machines" has been successfully selected as one of the seventeen German DFG priority programs which start in 2017[10]; "Argument Mining" is a British EPSRC-funded project implemented by the Centre for Argument Technologies in collaboration with IBM[11]; and "MIning and REasoning with legal text" is the EU H2020 Research and Innovation Staff Exchange project MIREL, in which mining legal arguments is one of the main tasks[12].

1.2 Explosion in the amount of data

Argument mining is an area of text mining which aims to develop techniques and methods for automated extraction of a variety of valuable information from large datasets of texts in natural language. The increase in the importance of text mining is the result of the challenge posed by the explosion of data available on the Internet. While having a vast amount of information is unquestionably of value, such resources become less useful, or even useless, if we cannot process the data efficiently and quickly enough. For example, customer feedback can be

[8] http://www.i3s.unice.fr/~villata/tutorialIJCAI2016.html
[9] http://acl2016tutorial.arg.tech/
[10] See e.g. http://www.eurekalert.org/pub_releases/2016-03/df-de1032416.php
[11] http://arg.tech/argument-mining
[12] http://mirelproject.eu/

an important source of information for companies such as Amazon or eBay, but manual analysis of this feedback will become ineffective if it is posted on the company's website faster than it can be processed by human analysts. This is often referred to as the Big Data problem. Text mining aims to develop robust tools and techniques to address this challenge and speed up the process of processing, interpreting and making sense of the large amount of natural language data.

Text mining focuses on several valuable types of information to be mined from natural language texts. Sentiment analysis aims to extract people's attitudes (positive, neutral or negative) towards persons, institutions, events and products, e.g. U.S. presidential candidates. It is applied in stock markets to quickly process vast resources, such as news and social media, to extract information about market trends and predict changes in stock prices. Opinion mining aims to mine opinions about persons, institutions, events and products, e.g. the opinion that the UK economy will be stronger without contributing money to the EU budget, or the contrasting opinion that the UK economy will be weakened without access to the common EU market. Its main commercial application is media analysis, which monitors various media to identify people's reactions to new products or political candidates. Finally, argument mining can be used to recognise not only *what* attitudes and opinions people hold, but also *why* they hold them. For example, people may think that the UK economy will be weakened without access to the common market because the EU will introduce high taxes for goods coming from the UK. The interest of various industries in this technology is manifested in the involvement of companies in a number of academic projects, as well as the development of techniques for argument mining in industry, such as IBM's Watson Debater, which searches for supporting and conflicting arguments on a given issue in Wikipedia articles[13]. One of the reasons for the huge upswing in interest in this area lies in the maturing technology and available returns. Opinion mining has transformed the way market research and PR are carried out, whilst sentiment analysis has had an even greater impact in predicting financial markets. Argument mining is the natural evolution of these technologies, providing a step change in the level of detail available – and this is why major organisations such as IBM are so interested in the technology.

1.3 Objects to be extracted in argument mining

Although argument mining always aims to extract arguments, the concept of "argument" itself may vary from one approach to another. This is the result of different applications, domains and theoretical foundations selected by different researchers. Two basic types of arguments are studied in the literature: supporting arguments and conflicting arguments.

In this section, we present an example of specification of these two notions in order to give the reader an intuitive understanding of the approach to argu-

[13]See e.g. `http://arg.tech/ibmdebater`

mentation in argument mining. Nevertheless, it is important to bear in mind that they can be defined in different ways in different argument mining work (see also subsection 2.2 for a few other examples of how these notions can be defined). This variety is important, as it allows the research to be adapted to different research tasks and applications. For example, in some cases it might be sufficient to identify only the two basic categories of "premise" and "conclusion", while for other research goals it might be necessary to recognise finer structures of supporting arguments, such as linked arguments (all the premises jointly support the conclusion), convergent arguments (the premises independently support the conclusion) and serial arguments (a chain of arguments). This variety, however, does have a drawback, as an analysis of argument structures (either manual or automatic) using one conceptualisation of argumentation might not be comparable to or reusable by another approach. To address this problem, the Argument Interchange Format (AIF) (see e.g. [Rahwan *et al.*, 2007; Chesnevar *et al.*, 2006; Bex *et al.*, 2012; Bex *et al.*, 2013; Budon *et al.*, 2014]) provides a standard for argument representation by capturing common argument structures shared by different approaches. An argument is here represented as a directed graph with two basic types of nodes: information – I-Nodes – and instances of schemes – S-Nodes, with sub-types of S-Nodes representing the application of rules of inference – RA-Nodes – and rules of conflict – CA-Nodes.

A *supporting argument* (also called an inference, pro-argument or argumentation) can be defined as a relation between propositions where p is used to support q[14], when p provides a reason to accept q. In other words, p can be used as a reply to the question "Why q?" ("Because p"). The concept of supporting argument is related to the traditional notion of (non-deductive) reasoning (see e.g. [Groarke, 1996; van Eemeren *et al.*, 2014]). In example (1), taken from the Regulation Room Divisiveness corpus[15], the annotator has interpreted the post by user sbarb95(272) as an argument in which the claim "I wonder if delays of 30 minutes would actually affect passenger behavior" is supported by the user's personal experience "In my experience it usually takes about 30 minutes to get to major airports".

(1) sbarb95(272): *(...) I wonder if delays of 30 minutes would actually affect passenger behavior. In my experience it usually takes about 30 minutes to get to major airports.*

In the Argument Interchange Format, the supporting argument is modelled as a graph consisting of two information nodes (in this example, a node repre-

[14]Note that there can be more than one support for q. In this case, we would have p_1, p_1, ..., p_n.

[15]In the US, online deliberative democracy or e-rulemaking, e.g. RegulationRoom (`http://RegulationRoom.org`), has been introduced as a multi-step process of social media outreach that federal agencies use to consult with citizens on new regulations on health and safety, finance, and other complex topics (see [Lawrence *et al.*, 2017] for the description of the corpus available at `http://arg.tech/rrd`).

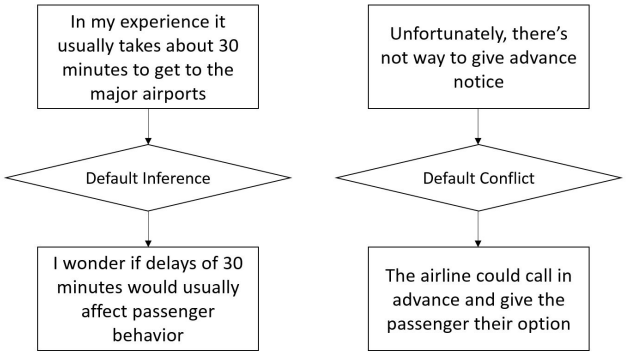

Figure 1. Supporting and conflicting arguments visualised according to the Argument Interchange Format (AIF) stored in the AIFdb database in the Regulation Room Divisiveness corpus, fragment of maps #4900 and #4891

senting the proposition "I wonder if delays of 30 minutes would actually affect passenger behavior" and a node representing "In my experience it usually takes about 30 minutes to get to major airports"); one node (RA-Node) representing the relation of support, called *Default Inference*; and an arrow representing the direction of reasoning – from premise(s) to conclusion (see the argument structure on the left side of Figure 1).

Supporting arguments are sometimes treated as a general, natural-language expression of support, even if actual reasons to accept a proposition are not provided, e.g. "Peanuts should be served on the plane" or "Peanuts are served on board on AirAlbatross flights". This broad notion is very close to the concept of support relation used in bipolar abstract argumentation [Cayrol and Lagasquie-Schiex, 2010].

A *conflicting argument* (also called an attack, con-argument or counter-argumentation) can be characterised as a relation between two propositions in which one proposition is used in order to provide an incompatible alternative to another proposition. Speakers use conflicting propositions to attack another speaker's claims with counter-claims. The concept of conflicting argument is close to the concept of attack relation used in abstract argumentation frameworks [Dung, 1995]. In a fragment of the discussion on the RegulationRoom platform (2), the analyst identified a conflicting argument between two forum users, AK traveler(287) and SofieM(947), in which "Unfortunately, there's no way to give advance notice" is used to attack the claim "The airline could call in advance and give the passenger their options".

(2) a. AK traveler(287): *(...) The airline could call in advance and give the passenger their options (...)*
b. SofieM(947): *(...) Unfortunately, there's no way to give advance notice.*

In AIF graphs, the conflicting argument is represented by a CA-node called *Default Conflict* (see the argument structure on the right side of Figure 1).

The conflict does not have to be restricted to the logical contradiction of p and not-p, but can be treated as a more general, natural-language expression of opposition, e.g. "Peanuts should be served on the plane" and "Peanuts cause allergic reactions". These two more general relations are used, for example, to identify conflicting and supporting viewpoints in online debates [Cabrio and Villata, 2012a].

Some approaches to formal argumentation focus only on one type of argument structure, such as the original abstract argumentation framework [Dung, 1995], while some work uses both supporting and conflicting arguments, such as ASPIC+ [Prakken, 2010].

2 Pipeline of natural language processing techniques applied to argument mining

In this section, we describe a pipeline which is typically used for automatic mining of arguments. The pipeline comprises two parts, linguistic and computational, which can be viewed as analogues (see Figure 2). The linguistic part aims to develop large corpora, which are datasets of manually annotated (analysed) argument data, evaluated by measuring the level of inter-annotator agreement. The computational part of the argument mining pipeline aims to develop grammars (structural approach) and classifiers (statistical approach) to create automatically annotated corpora of arguments, and the performance of the system is then evaluated by measures such as accuracy or F_1- score. The ultimate goal of the pipeline is to process real arguments in natural language texts (such as arguments formulated on Wikipedia) so that the output is only the information that is valuable to us, i.e. structured argument data. In subsequent subsections we briefly describe how each step of this process is typically developed and give a few examples of specific work and approaches, to help the reader to better understand what types of methods, techniques and results are likely to be found in papers in this area.

2.1 Databases of texts in natural language

The argument mining pipeline starts with the task of collecting large resources of natural language texts (see "large resources of NL text" box in Figure 2), which then can be used for training and testing of argument mining algorithms. Texts sometimes need to be pre-processed in order to obtain a common and processable format.

The size of the datasets varies from project to project, but in the case of approaches which apply machine learning techniques (see Section 2.5), the dataset should be relatively large, because the algorithms learn rules for argument mining using statistical correlations. In the case of structural approaches, such rules (a grammar) are developed by a linguist and thus the dataset can be smaller, as its size is complemented by the linguist's expertise and knowledge about

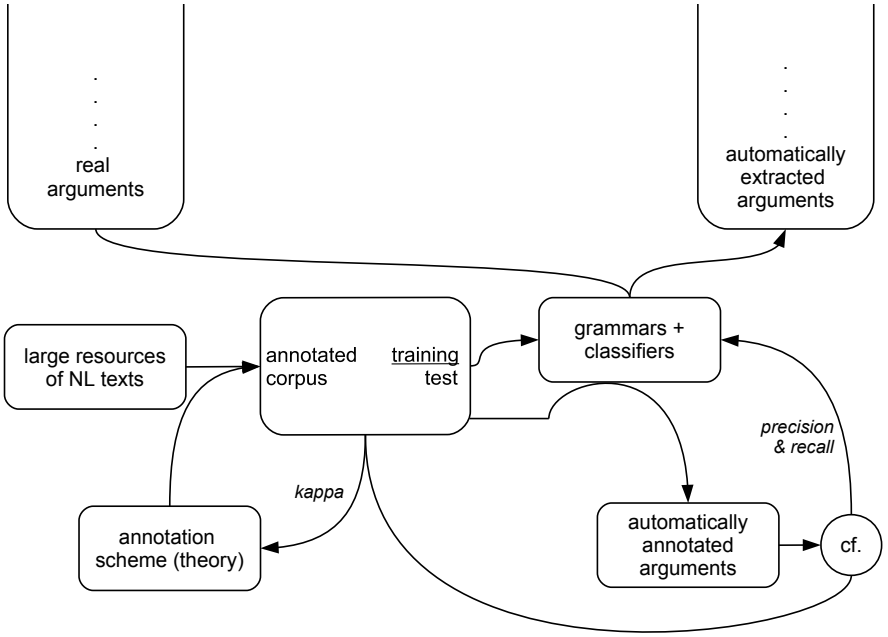

Figure 2. A pipeline of natural language processing techniques applied to argument mining (all steps in the pipeline will be explained in subsequent subsections of this section)

language.

For example, in the case of the statistical approach, Palau and Moens used a dataset consisting of 92,190 words (2,571 sentences) in 47 documents from the European Court of Human Rights [Palau and Moens, 2009], and Habernal and Gurevych collected a database comprising 90,000 words in 340 documents of user-generated web discourse [Habernal and Gurevych, 2016a]. In the case of the structural approach, Garcia-Villalba and Saint-Dizier used 21,500 words in 50 texts as a test corpus [Villalba and Saint-Dizier, 2012], and Budzynska, Janier, Reed and Saint-Dizier collected a dataset consisting of 24,000 words [Budzynska *et al.*, 2016].

Typically, the task of argument mining is narrowed down to a specific type of discourse (genre), since algorithms use the linguistic surface for argument recognition with little or no knowledge about the world, the discourse context or the deeper pragmatic level of a text. In other words, the way people express argumentation in language depends on the type of discourse they are engaged in; for example, we can expect a lawyer in court to argue in a different linguistic style than a politician during a presidential debate or a scientist at a conference.

Genres that have been studied include legal texts [Moens *et al.*, 2007; Reed *et al.*, 2008; Palau and Moens, 2009; Ashley and Walker, 2013; Wyner *et al.*, 2016]; mediation [Janier *et al.*, 2015; Janier and Reed, 2015]; scientific papers [Teufel *et al.*, 1999; Teufel *et al.*, 2009; Kirschner *et al.*, 2015]; student essays [Nguyen and Litman, 2015; Nguyen and Litman, 2016]; online comments [Villalba and Saint-Dizier, 2012; Park and Cardie, 2014b; Park and Cardie, 2014a; Habernal *et al.*, 2014; Wachsmuth *et al.*, 2015; Konat *et al.*, 2016]; political debates [Hirst *et al.*, 2014; Naderi and Hirst, 2015; Duthie *et al.*, 2016b]; philosophical texts [Lawrence *et al.*, 2014]; technical texts [Marco *et al.*, 2006; Saint-Dizier, 2014]; moral debate on the radio [Budzynska *et al.*, 2014b]; online debates [Swanson *et al.*, 2015; Walker *et al.*, 2012; Cabrio and Villata, 2014; Sridhar *et al.*, 2015; Boltuzic and Snajder, 2016; Habernal and Gurevych, 2016c]; persuasive essays [Stab and Gurevych, 2014; Ghosh *et al.*, 2016]; and Wikipedia articles [Aharoni *et al.*, 2014; Levy *et al.*, 2014; Lippi and Torroni, 2015b; Roitman *et al.*, 2016]. The choice of genre is the researcher's subjective decision, but is typically motivated by the presumed likelihood of finding a large amount of non-sparse (not rare) argument data in a given type of discourse (in other words, how argumentative a given genre seems to be). Sometimes, if a project is run in cooperation with an industrial partner, the choice also depends on commercial applications.

2.2 Theories & annotation schemes

The next step in the argument mining pipeline consists of choosing a model of argumentation, which is then used to develop an annotation scheme for analysing argumentation in natural language texts. An annotation scheme is a set of labels (tags) which defines argumentation and its aspects, as well as which annotators (analysts) will be used to structure the dataset.

This task is important but also challenging, as a faulty conceptualisation of the phenomenon of argumentation results in erroneous and inconsistent annotations (analysts tend to make more errors if they do not understand the annotation scheme, and different analysts will make different decisions; see Section 2.4), which will then have negative consequences for the performance of the automation process (see Section 2.5). In other words, if the human analyst struggles with understanding a given annotation scheme, then we should not expect the algorithm to be able to learn it efficiently. As a result, the extraction of arguments will not be accurate (see Section 2.7), as the algorithm is unable to detect the correct patterns between the annotated tags and the linguistic surface of the annotated texts.

The literature contains a variety of annotation schemes which aim to achieve a balance between efficiency (simpler schemes will be quicker and easier to annotate) and adequacy (more specific sets of labels will be better tailored to describing given aspects of argumentation or a given genre). In one of the first works in argument mining, [Moens *et al.*, 2007; Palau and Moens, 2009], Palau and Moens choose a basic, intuitive conceptualisation of argument structure:

- Premise: statement which provides support
- Conclusion: statement which is supported
- Argument: a full structure comprising the premises and conclusion

Such a set of labels is unquestionably the most natural and obvious choice to start with for the purpose of mining argumentation. It is also a general, high-level description and can be reused for any other genre. The only limitation that must be accounted for is that the property of being a premise and the property of being a conclusion are in fact relations, i.e. a premise cannot merely provide support – it always provides support for something. For instance, the sentence "home ownership may not be a basic human right" uttered in isolation is just a statement (an assertion) of someone's belief; it becomes a premise only if it is a response to or a continuation of another utterance, such as "more successful countries, like Switzerland and Germany, prefer to rent". Thus, we need another way to account for this relational context, so that the algorithm has enough information to learn argument structure properly.

In her work on Argumentative Zoning [Teufel *et al.*, 1999; Teufel *et al.*, 2009], Teufel uses a more complex set of labels specifically tailored for mining argumentation in scientific texts:

- Background: general scientific background
- Other: neutral descriptions of other people's work
- Own: neutral descriptions of our own, new work
- Aim: statements of the particular aim of the current paper
- Textual: statements characterising the textual organisation of the current paper (e.g. "In chapter 1, we introduce...")
- Contrast: contrastive or comparative statements about other work; explicit mention of weaknesses of other work
- Basis: statements that our own work is based on other work

The advantage of an annotation scheme tailored to a specific genre is that specific categories are able to capture more detailed information about typical language constructions in a given type of discourse, and thus to enhance the development of rules or the learning process. This advantage, however, must be balanced against the limited reusablity of such an annotation scheme for other genres.

Peldszus and Stede [Peldszus and Stede, 2013] introduce an annotation scheme drawing on different ideas from the literature and their own practical experience analysing texts in the Potsdam Commentary Corpus [Stede,

2004]. The schema follows Freeman's idea of using the moves of the proponent and challenger in a basic dialectical situation as a model of argument structure [Freeman, 1991; Freeman, 2011]. The authors define an argument as a non-empty set of premises supporting some conclusion, using the term "argument" not for premises, but for the complex of one or more premises put forward in favour of a claim. Premises and conclusions are propositions expressed in text segments. If an argument involves multiple premises that jointly support the conclusion, a linked structure is identified (i.e. none of the linked premises would be able to support the conclusion on its own). In the basic dialectical situation, a linked structure is induced by the challenger's question as to why a premise is relevant to the claim. The proponent then answers by presenting another premise explicating the connection. Building a linked structure is conceived as completing an argument. In this scheme, the label "argumentation" refers to the structure that emerges when multiple arguments are related to each other and form larger complexes. The manner in which arguments combine into larger complexes can be generally described as either supporting, attacking or counter-attacking. More precisely, the scheme considers five kinds of support:

- basic argument
- linked support
- multiple support
- serial support
- example support

four kinds of attacks on the proponent's argument by the challenger:

- rebut a conclusion
- rebut a premise
- undercut an argument
- support a rebutter

and four counter-attacks on the challenger's attack by the proponent:

- rebut a rebutter
- rebut an undercutter
- undercut a rebutter
- undercut an undercutter

Stab and Gurevych [Stab and Gurevych, 2014] propose an annotation scheme whose goal is to model argument components as well as the argumentative relations that constitute the argumentative discourse structure in persuasive essays[16]. Essays of this kind exhibit a common structure: the introduction typically includes a major claim that expresses the author's stance on the topic, and the major claim is supported or attacked by arguments (composed of premises and a claim) covering certain aspects of the stance in subsequent paragraphs. The major claim is the central component of an argument. It is a controversial statement that is either true or false and should not be accepted by the reader without additional support. The premise underpins the validity of the claim and is a reason given by the author to persuade the reader. Argumentative relations model the discourse structure of arguments and indicate which premises belong to a given claim. In this scheme, two directed relations between argument components are annotated: the support relation and the attack relation. Both relations can hold between a premise and another premise, a premise and a claim, or a claim and a major claim. An argumentative relation between two components indicates that the source component is a reason for or refutation of the target component. The annotation guidelines proposed in [Stab and Gurevych, 2014] consist of the following steps:

- Topic and stance identification: before starting the annotation process, annotators identify the topic and the author's stance by reading the entire essay.
- Annotation of argument components: the major claim is identified in either the introduction or the conclusion of the essay. Subsequently, annotators identify the claims and premises in each paragraph.
- Annotation of argumentative relations: finally, the claims and premises are linked within each paragraph, and the claims are linked to the major claim with either a support relation or an attack relation.

Finally, an annotation scheme which considers the broad dialogical context of argumentation can be found in [Budzynska *et al.*, 2014b]. Building upon Inference Anchoring Theory (IAT) [Budzynska and Reed, 2011] as a theoretical foundation, they propose to extend the set of tags for supporting and conflicting arguments with dialogical and illocutionary structures (the latter are modelled as illocutionary forces introduced in [Austin, 1962; Searle, 1969; Searle and Vanderveken, 1985]). The illocutionary structures are described by two groups of tags:

- Tags associated with a speaker's individual moves in the dialogue:
  - Asserting: The speaker S *asserts* p to communicate his opinion on p.

[16]The corpus and annotated guidelines are available at `https://www.ukp.tu-darmstadt.de/data/argumentation-mining`.

  - Questioning: S *questions whether p* when S formulates p as an interrogative sentence of the form "Is/Isn't p the case?" Three categories of questioning are distinguished: Pure Questioning (PQ), Assertive Questioning (AQ), and Rhetorical Questioning (RQ). In the case of PQ, S is asking for the hearer H's opinion on p: whether H believes, disbelieves, or has no opinion on p. AQ and RQ, in contrast, carry some degree of assertive force and thus can be a part of the argumentation itself. For AQ, S not only seeks H's opinion on p, but also indirectly declares publicly his own opinion on p. For RQ, S is grammatically stating a question, but in fact he conveys that he does (or does not) believe p.
  - Challenging: When S *challenges p*, S declares that he is seeking (asking about) the grounds for H's opinion on p. Challenges are a dialogical mechanism for triggering argumentation. Like questions, challenges form a continuum from Pure Challenging (PCh) to Assertive Challenging (ACh) to Rhetorical Challenging (RCh).
  - Popular Conceding: Through *popular conceding*, S communicates some sort of general knowledge which is assumed to be obvious and as such does not need to be defended (does not place a burden of proof on S).

- Tags associated with the interactions between speaker(s)' moves in the dialogue:

  - Agreeing: Agreeing is used to express a positive reaction, i.e. when the speaker S declares that he shares the opinion of the hearer. This can take the form of signalling such as "Yes", "Indeed", "Most definitely" or "Sure", but may also be a complete sentence.
  - Disagreeing: Disagreeing is used to express a negative reaction, i.e. when S declares that he does not share H's opinion. This can take the form of utterances which have similar meaning to "No" (e.g. "I'm not saying that", "Actually, that's not correct", "Definitely not" or "No it's not"), or it can be an utterance with complete propositional content.
  - Arguing: S is *arguing* when he defends a standpoint. This illocution is sometimes signalled by linguistic cues such as "therefore", "since" or "because", but these indicators rarely occur in natural spoken language.

The complexity of the annotation scheme allows more aspects of natural communication to be captured (in the case of this project, both argument and dialogical structures, as well as the relations between them), but on the other hand, the annotation process becomes a more time-consuming task and requires well trained and experienced annotators.

2.3 Manual annotation & corpora

Once we collect a database of texts for our genre and select or develop an annotation scheme, we need to combine these two together, i.e. an annotator manually assigns a label from the scheme to a segment of the text. The resulting database of tagged texts is called a corpus.

More specifically, the annotation process starts with segmentation (splitting) of the text into elementary discourse units (EDUs; see e.g. [Lawrence *et al.*, 2014]) or, more precisely, argumentative discourse units (ADUs; see e.g. [Peldszus and Stede, 2013; Ghosh *et al.*, 2014]). ADUs present an additional challenge for the segmentation, since they require the analyst not only to find spans of text which are minimal meaningful building blocks of a discourse (EDU)[17], but also to recognise whether or not they carry an argumentative function (whether or not an EDU is an ADU). Next, the annotator assigns a label from the scheme to the ADU following the guidelines, i.e. the definitions of the labels.

Annotators typically use software tools, such as the `arggraph` DTD[18], the RSTTool[19] and the Glozz annotation tool[20], to help them assign labels from the annotation scheme to ADUs directly in a code. The method often uses the XML format of tagging text. Take this example of an ADU:

(3) Ruth Levitas: *No parent in the family is in work since we have a huge problem with unemployment.*

This ADU can be represented by the following string of XML code:

```
<argumentation><conclusion>No parent in the family is in
 work</conclusion> since <premise>we have a huge problem with
 unemployment</premise></argumentation>
```

Another approach would be to exploit a visualisation software tool such as GraPAT[21] [Sonntag and Stede, 2014], which is a graph-based, web-based tool suited for annotation of sentiment and argumentation structure. The tool provides a graph structure annotation for the selected text. Certain parts of the text can be connected to nodes in the graph, and those nodes can be connected via the relations of the selected annotation scheme. Edges also have attributes to describe them. Figure 3 shows an example of annotation of argumentation structure using the GraPAT tool.

[17] There is no general agreement on one standard definition of EDU.
[18] `https://github.com/peldszus/arg-microtexts/blob/master/corpus/arggraph.dtd`
[19] `http://www.wagsoft.com/RSTTool/`
[20] `http://www.glozz.org`
[21] Available at: `http://angcl.ling.uni-potsdam.de/resources/grapat.html`

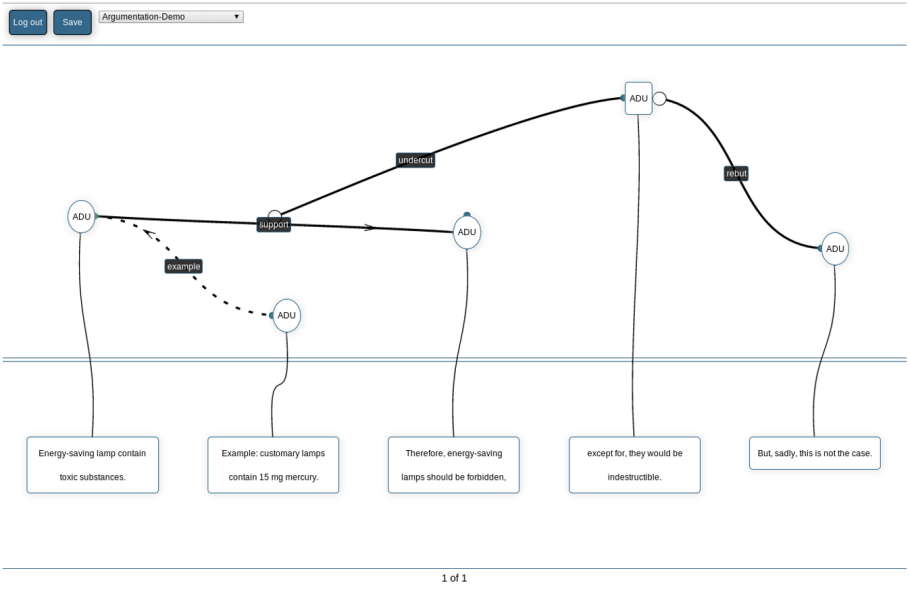

Figure 3. Annotation of an argumentation structure in GraPAT

To capture more detailed information about the dialogical context of argumentation, the Online Visualisation of Arguments tool (OVA+)[22] [Janier *et al.*, 2014] can be used (see Figure 4). In [Budzynska *et al.*, 2016], this tool was used to analyse argument structures in broadcast debates from the BBC Radio 4 programme *Moral Maze*. OVA+ allows the annotator to represent the communicative intention of arguing in Example (3) as a relation between the *inference relation* (Default Inference in Figure 5) and the *dialogical relation* (Default Transition in Figure 5)[23]. In other words, we can say that the inference is the content of arguing, while the transition between moves in the dialogue anchors this intention. In a similar manner, we can annotate an illocution of disagreeing in order to link a conflict (a conflicting argument) between propositions with the dialogical relation of transition (see Figure 6).

Finally, the annotated data must be stored as a corpus ([Reed, 2006] is one of the first corpora of analysed arguments). For example, the IBM Debating

[22]Available at: `http://ova.arg-tech.org`

[23]More precisely, the graph illustrates that the conclusion (top left node in Figure 5) is introduced by Levitas' first utterance (top right node) via the communicative intention of asserting (middle top node); the premise (bottom left node) is introduced by her second utterance (bottom right node), also via the communicative intention of asserting (middle bottom node), while the relation of inference (*Default Inference* node in Figure 5) is triggered by the interaction between these two moves (*Default Transition* node) via the communicative intention of arguing.

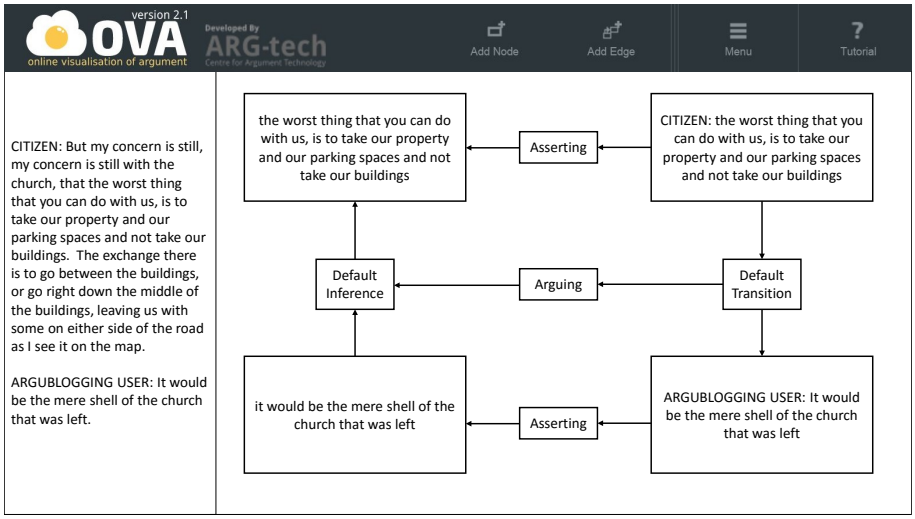

Figure 4. Annotation of an argumentation structure in OVA+. The panel on the left side shows the raw text to be annotated, and the panel on the right shows the visualisation of the annotation, i.e. argument and dialogue structures represented as a directed graph.

Technologies corpus[24] contains three different datasets: a dataset for automatic detection of claims and evidence in the context of controversial topics (1,392 labelled claims for 33 different topics) [Aharoni *et al.*, 2014]; its extended version (2,294 labelled claims and 4,690 labelled evidence for 58 different topics) [Rinott *et al.*, 2015]; and a dataset of multi-word term relatedness (term-relatedness values for 9,856 pairs of terms) [Levy *et al.*, 2015].

Another important resource is the Internet Argument Corpus (IAC), which provides analyses of political debate on Internet forums. It consists of 11,000 discussions and 390,000 posts annotated for characteristics such as topic, stance, degree of agreement, sarcasm, and nastiness, among others [Walker *et al.*, 2012]. It does not include argument components but includes several properties of arguments, and it can be used for argument attribution tasks. The UKPConvArg1[25] corpus is another recently released dataset, composed of 16,000 pairs of arguments on 32 topics annotated with the relation "A is more convincing than B" [Habernal and Gurevych, 2016c].

The application of labels and Inference Anchoring Theory to the dataset of transcripts of the BBC Radio 4 programme *Moral Maze* (the labels were described in Section 2.2) resulted in the MM2012 corpus [Budzynska *et al.*, 2014b] (see Figure 8 for an example[26]). Figures 5 and 6 depict fragments of

[24] `http://researchweb.watson.ibm.com/haifa/dept/vst/mlta_data.shtml`
[25] `https://github.com/UKPLab/acl2016-convincing-arguments`
[26] Available at: `http://corpora.aifdb.org/mm2012`

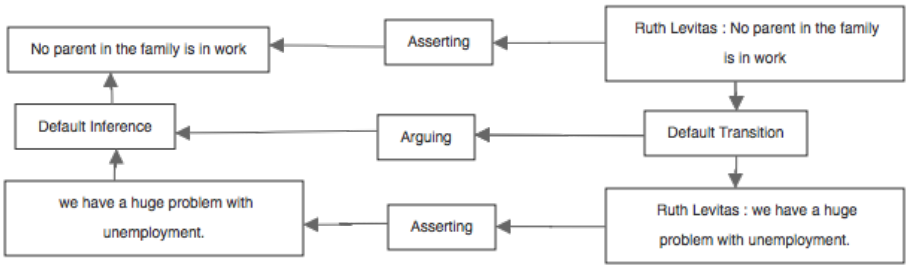

Figure 5. Annotation of *Arguing* and *Inference* using the IAT annotation scheme and the OVA+ software tool, stored in the AIFdb database in the Moral Maze 2012 corpus (fragment of map #6306)

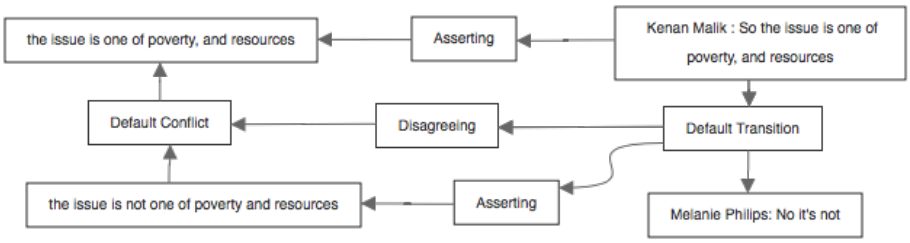

Figure 6. Annotation of *Disagreeing* and *Conflict* using the IAT annotation scheme and the OVA+ software tool, stored in the AIFdb database in the Moral Maze 2012 corpus (fragment of map #6331)

maps representing structure of argument data[27]. Table 1 shows the frequency of tags associated with the interactions between moves of speaker(s) in the dialogue: 1,191 such tags were used to annotate the MM2012 corpus, and arguing was the most frequently occurring tag (72% of cases).

As manual annotation is a highly time-consuming task, sharing and reusing annotated data is of real value[28]. This is the aim of the freely accessible

[27] Available at: `http://corpora.aifdb.org/mm2012`

[28] See for example the "Unshared task" session at the 3rd Argument Mining workshop or " Unshared untask" sessions at Dagstuhl's *Natural Language Argumentation* Seminar.

IC type	**Occurrences**	**Kappa κ**
Agreeing	119 (10%)	
Disagreeing	219 (18%)	
Arguing	853 (72%)	
TOTAL	**1,191** (100%)	**.76**

Table 1. The distribution of illocutionary connections anchored in transitions in the MM2012 corpus

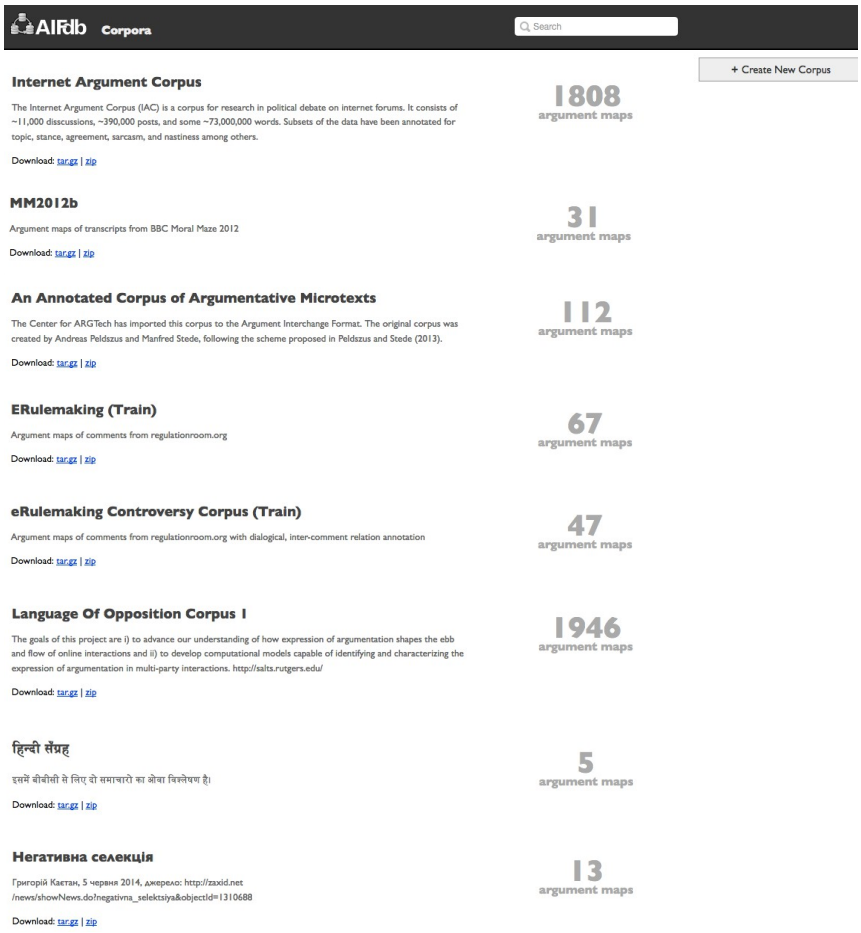

Figure 7. Freely available AIFdb corpora

database AIFdb [Lawrence *et al.*, 2012][29], which hosts multiple corpora [Lawrence *et al.*, 2015][30]. The key advantage of AIFdb is that it uses a standard for argument representation – the Argument Interchange Format (AIF) [Rahwan *et al.*, 2007; Chesnevar *et al.*, 2006]. The database stores corpora which were either originally annotated according to this format, such as the MM2012 corpus described above, or imported to the AIFdb, such as the Internet Argument Corpus developed in Santa Cruz [Walker *et al.*, 2012], the Microtext corpus developed in Potsdam [Peldszus and Stede, 2015], the Erulemaking corpus from Cornell [Park and Cardie, 2014b], and the Language of Opposition corpus de-

[29] `http://aifdb.org`
[30] `http://corpora.aifdb.org/`

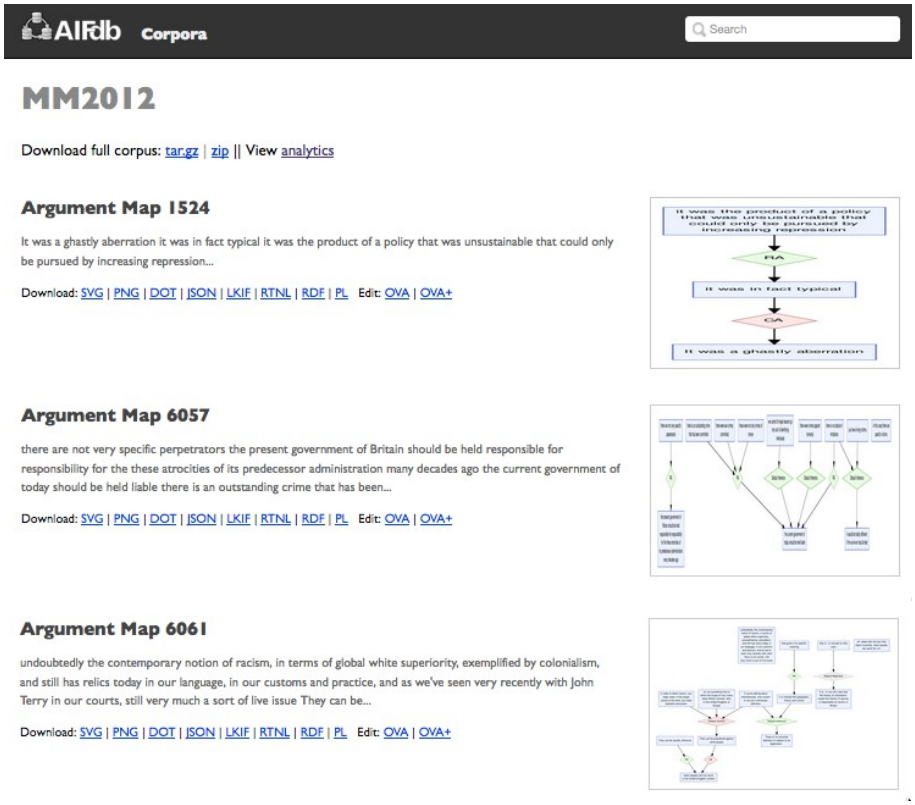

Figure 8. The Moral Maze 2012 corpus created, stored and managed on the AIFdb corpora platform

veloped at Rutgers and Columbia [Ghosh *et al.*, 2014; Wacholder *et al.*, 2014][31] (see Figure 7). The AIF standard allows for the expression of a variety of argument aspects shared by many approaches to argumentation, and thus for the reusability of the resources analysed according to this format. The AIFdb corpora provide the service of creating (see the button in the top right hand corner in Figure 7), storing and managing the annotator's own corpus and allow the stored data to be reused by other researchers and projects. Currently the AIFdb database has 300-500 unique users per month and stores 1,500,000 words and 50,000 annotated arguments in 15 languages (statistics obtained in July 2016).

2.4 Evaluation of the manual annotation step

The last step of the linguistic part of the argument mining pipeline is the evaluation of the corpus. This task aims to reduce the propagation of errors

[31] The data and coding are available at: `http://salts.rutgers.edu/`.

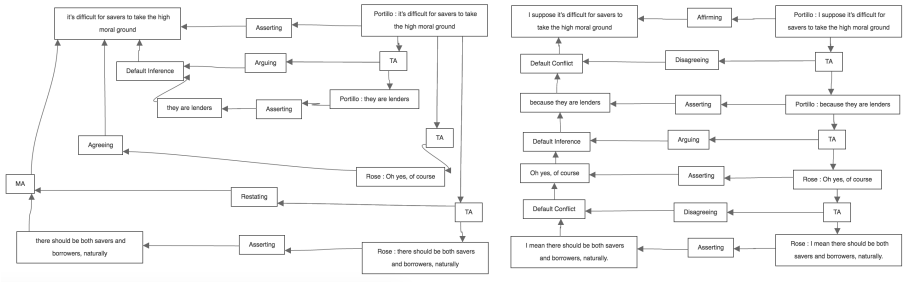

Figure 9. An evaluation of the quality of manual annotation using inter-annotator agreement: the argument map on the left analysed by one annotator is compared with the argument map on the right analysed by another annotator.

from the manual annotation to the automatic annotation in order to improve the performance of the system.

The quality of annotation is evaluated using a measure of inter-annotator agreement, i.e. by comparing analyses of the same texts done by different annotators. Most typically, the comparison is calculated for a subset of the corpus – one annotator analyses the full database, while the other analyses just a subset of this database. For example, we can look for matches between the map on the left hand side and the map on the right in Figure 9, counting how many nodes and relations were analysed by both annotators in the same way.

The two comparison measures used most often are agreement and kappa κ. Agreement is a simple proportion (percentage) of matches. This measure, however, does not take into account the possibility of random matches, as if the annotators were tossing a coin and then assigning labels according to the result. For this reason the more informative kappa measure was introduced. Its most popular version is Cohen's kappa [Cohen, 1960], which shows the agreement between two annotators who each classify N items (e.g. ADUs) into C mutually exclusive categories (tags). The equation for κ is as follows:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where $Pr(a)$ is the relative observed agreement among raters and $Pr(e)$ is the hypothetical probability of chance agreement. The work [Landis and Koch, 1977] proposes a scale for κ results to help to interpret the level of agreement. According to these authors, 0.41-0.60 represents moderate agreement, 0.61-0.80 is substantial agreement, and 0.81-1 is almost perfect agreement. It should be noted, however, that this scale is not widely accepted.

Most recently, Duthie, Lawrence, Budzynska and Reed proposed a new met-

Method	Overall Score
Cohen's κ	0.44
CASS-κ	0.59
$F1$	0.66
CASS-$F1$	0.74

Table 2. Scores are provided for Cohen's kappa and $F1$ score, for both segmentation and structure, and their CASS equivalence.

ric [Duthie *et al.*, 2016a]. The Combined Argument Similarity Score (CASS) technique aims to account for the problem of over-penalising or double-penalising small differences between two manual annotations or between manual and automatic annotation. This technique looks to break down the argument structure into three components: segmentation, argumentative relations and dialogical relations. Segmentation calculations are particularly important when performing argument mining on unaltered natural language text. Arguments do not always span full sentences and automatic solutions may miss some tokens, which can then have a knock-on effect on the argumentative or dialogical structure, with the same ADU segmented in a slightly different way by two different analysts and κ penalising for this twice. CASS offers the Segmentation Similarity metric [Fournier and Inkpen, 2012] to give a score purely for segmentation, which does not penalise as heavily for near-misses. Further, it uses the Levenshtein distance [Levenshtein, 1966] together with word positions to match text spans and compare the relations between them in order to work out scores for both argumentative structure and dialogical structure. CASS then takes the mean of the argumentative and dialogical scores and the harmonic mean to calculate the CASS score. The resulting differences for standard metrics and the new CASS metrics can be seen in Table 2.

CASS is implemented and made available as part of Argument Analytics [Lawrence *et al.*, 2016], a suite of techniques which provides an interpretation of and insight into large-scale argument data stored in AIFdb corpora for both specialist and general audiences. Currently, it has modules available for viewing the following: simple statistical data, which give both an overview of the argument structure and frequencies of patterns such as argumentation schemes; comparative data, which can be quantified by a range of measures (including the CASS metric) describing the similarity of two annotations; dialogical data showing the behaviour of participants of the dialogue; and real-time data allowing for the graphical representation of an argument structure developing over time.

In the eRulemaking corpus [Park and Cardie, 2014a], inter-annotator agreement was measured on 30% of the data for two annotators, resulting in a Cohen's κ of 0.73; in the Moral Maze corpus [Budzynska *et al.*, 2014b], Cohen's kappa for two annotators and three types of illoctionary connections (arguing, agreeing and disagreeing) was $\kappa = .76$ (see also Table 1); in the Persuasive

Essays corpus [Stab and Gurevych, 2014], inter-annotator agreement was measured on 90 persuasive essays for three annotators, resulting in a Krippendorff's inter-rater agreement of $\alpha = 0.81$[32]; and in the argumentative microtexts corpus [Peldszus and Stede, 2015], three annotators achieved an agreement of Fleiss' $\kappa = 0.83$[33] for the full task.

2.5 NLP techniques

The next step initiates the computational part of the argument mining pipeline (see the "grammars + classifiers" box in Figure 2). In principle, there are two basic styles of automation (in practice, they in fact share some techniques, to the extent that they sometimes form a hybrid approach): the structural approach, i.e. grammars (hand-coded sets of rules) and the statistical approach, i.e. machine learning (general learning algorithms).

In the first case, the linguist looks through a selected fragment of a corpus and aims to find patterns in the expression of argumentation in natural language. For instance, in a given corpus it might be observed that arguments are linguistically signalled by words such as "because", "since" or "therefore", or that assertive questions typically start with "Isn't it..." or "Can we agree that...", etc. In such a case the linguist will formulate rules describing these patterns and add them to the grammar. Then, the grammar will be used by a system for automatic recognition of categories of argumentation and assertive questioning.

The statistical approach replaces a linguist with an algorithm. Like a human being, a system will also look for patterns, but this time on a larger sample of the training corpus. The programmer chooses what features the machine should process to check whether they coexist with the categories from the annotation scheme. Commonly chosen features include the following:

- unigrams: each word in the sentence
- bigrams: each pair of successive words
- text statistics: sentence length, average word length, number of punctuation marks
- punctuation: the sequence of punctuation marks present in the sentence
- keywords: words which are typically associated with a category, e.g. "but" and "however"for conflicting argument, or "because", "since" and "therefore" for supporting argument.

Both methods have their advantages. The structural approach exploits the expert's knowledge and experience with language analysis, which is helpful in

[32]Krippendorff's α is a statistical measure of the agreement achieved when coding a set of units of analysis in terms of the values of a variable.

[33]Fleiss' κ assesses the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items.

formulating correct patterns and, at the same time, to ignore random statistical correlations. On the other hand, a machine can easily spot new and infrequent patterns which are difficult for a linguist to spot due to the smaller sample size that can be analysed by a human being or to bias arising from prior knowledge and expectations. Algorithms are also able to quickly and easily update the learning process, while extending a grammar requires more effort. A lot of work in argument mining applies typical, "off-the-shelf" NLP methods and techniques, which in order to improve the performance of argument mining systems are adapted to the domain of argumentation by further enriching them with discourse indicators [Knott, 1996; Henkemans *et al.*, 2007], argumentation schemes [Feng and Hirst, 2011; Lawrence and Reed, 2016], frames (aspects of an issue) [Naderi and Hirst, 2015], dialogical context [Budzynska *et al.*, 2014a; Budzynska *et al.*, 2015], semantic context [Cabrio and Villata, 2012a], and combinations of various cues and techniques [Lawrence and Reed, 2015].

An example of the structural approach is the framework proposed by Garcia-Villalba and Saint-Dizier [2012], which aims to extract arguments in opinion texts. The authors discuss how automatic recognition of arguments can be implemented in the Dislog programming language on the <TextCoop> platform. The main idea of this work is that argument mining techniques make it possible to capture the underlying motivations consumers express in reviews, which provide more information than a basic attitude of the type "I do/don't like product A". Extraction of arguments associated with evaluative expressions provides a deeper understanding of consumer motivations and preferences. Evaluative expressions, more precisely, are accompanied by additional elements such as comments, elaborations, comparisons, illustrations, etc., which are forms of explanation. According to the authors, explanation is not a basic rhetorical relation, but a very generic construction which covers a large number of communication situations and which is realised in language by means of several types of rhetorical relations (e.g. elaboration, arguments or illustration). Garcia-Villalba and Saint-Dizier argue that rhetorical elements related to explanation behave as argument supports and make explicit the semantic and pragmatic function of the support: they justify, illustrate and develop the evaluative expression. The main hypothesis is that while elaborations, illustrations and other rhetorical relations related to explanation have no argumentative orientation, they acquire such orientation when combined with an evaluative expression. In the example given by the authors, a justification gives a reason for the evaluation expressed in the review: "The hotel is 2 stars [JUSTIFICATION due to the lack of bar and restaurant facilities]" can be classified as a justification whose general abstract schema is "X is Eval because of Fact*", where Eval denotes the evaluative expression and Fact* is a set of facts acting as justifications. The contrast relation introduces a statement which is somewhat symmetric but in opposition with the evaluation, e.g. "It was clean and comfortable, [CONTRAST but the carpet was in need of replacing]" whose schema is "X is Eval but B*", for which examples of linguistic markers include

"whereas", "but" and "while".

In order to process these relations, the authors use <TextCoop>, a platform designed for discourse analysis with a logic and linguistics perspective. Argument extraction rules correspond to the types of discourse structures that appear in consumer opinion texts. These rules have been adapted from a rule repository dedicated to explanation structures by developing additional lexical resources (terms expressing polarity, adverbs of intensity or domain verbs) and revising the structure of rules. For example, for the contrast relation, here are the six rules that have been defined:

Contrast ⇒ conn(opposition whe), gap(G), ponct(comma). /
conn(opposition_whe), gap(G), end. /
conn(opposition_how), gap(G), end.

where conn(opposition_whe): whereas, but whereas, but while, and conn(opposition_how): however.

The majority of the work in argument mining employs the statistical approach. For instance, binary supervised classifiers are trained to distinguish argumentative sentences from non-argumentative ones. Several standard machine learning algorithms have been exploited for argument mining tasks, such as Support Vector Machines (SVM), Logistic Regression, Naive Bayes classifiers, Decision Trees, and Random Forests. They are trained in a supervised setting, starting from a set of annotated examples (the training set of the corpus), and are then tested on unseen examples (the test set of the corpus). The annotated instances are provided to the system in the form of a feature vector, and each example is associated with the appropriate class. Concerning the choice of features, many of the existing approaches in the literature share almost the same set of features, including classical features for text representation (such as those mentioned above), sentiment-based features, and Bag-of-Words (BoW) representations, where the sentence is encoded as a vector of binary values over huge dictionaries. BoWs for bigrams and trigrams are also possible. Recent approaches to argument mining are also beginning to exploit deep neural architectures, such as Recurrent Neural Networks (RNN), a popular model in NLP tasks. In a nutshell, the idea behind RNN is to make use of sequential information, i.e. to perform the same task for every element of a sequence, with the output depending on the previous computations. The most commonly used type of RNNs are Long Short-Term Memory neural networks (LSTMs), which are better at capturing long-term dependencies than basic RNNs.

Among these, Lippi and Torroni [Lippi and Torroni, 2015b] present a framework to detect claims in unstructured corpora without necessarily resorting to contextual information. Their methodology is driven by the observation that argumentative sentences are often characterised by common rhetorical struc-

tures. As the structure of a sentence can be highly informative for argument detection, and in particular for identification of a claim, the authors choose constituency parse trees to represent such information. They therefore build a claim detection system based on an SVM classifier which aims at capturing similarities between parse trees through Tree Kernels, a method used to measure the similarity between two trees by evaluating the number of their common substructures.

Another recent statistical approach to argument mining has been presented by Habernal and Gurevych [Habernal and Gurevych, 2016b], who tackle a new task in computational argumentation, aiming to assess the qualitative properties of arguments in order to explain why one argument is more convincing than another. Based on a corpus of 26,000 annotated explanations written in natural language, two tasks are proposed on this dataset: prediction of the full label distribution and classification of the types of flaws in less convincing arguments. They define a framework composed of feature-rich SVM learners and Bidirectional LSTM neural networks with convolution.

In Section 3, we present an example of the statistical approach in more detail.

2.6 Automatically annotated data

A system developed in the NLP stage is then used to process raw, unannotated text in order to automatically extract arguments. The text is taken from the test corpus with all the tags removed. This step can be viewed as an automated equivalent to the manual annotation and corpus development described in Section 2.3. In other words, the NLP system creates an automatically annotated corpus.

In the work by Lippi and Torroni [Lippi and Torroni, 2015b] described in the previous section, automatically annotated data are visualised using colour highlighting (see Figure 10): the claims detected by their system are highlighted in red; false positives (detected by the system but not labelled as context-dependent claims for a given topic in the manually annotated IBM corpus) are blue; and false negatives (not detected by the system but labelled as positives in the manually annotated IBM corpus) are green.

Figure 11 shows another example of the output of a software tool. The <TextCoop> platform [Saint-Dizier, 2012], described above, produces automatic segmentation and annotation. The text is split into argumentative discourse units (ADUs) which contain minimal meaningful building blocks of a discourse with argumentative function. These propositional contents are presented as text in purple. Then, the system assigns illocutionary connections (text in green) to ADUs which are assertions, rhetorical questions (RQ), and so on, as well as polymorphic types to represent ambiguity (or underspecification), such as RQ-AQ, which means that an ADU can be anchored in the discourse via either rhetorical questioning or assertive questioning (AQ).

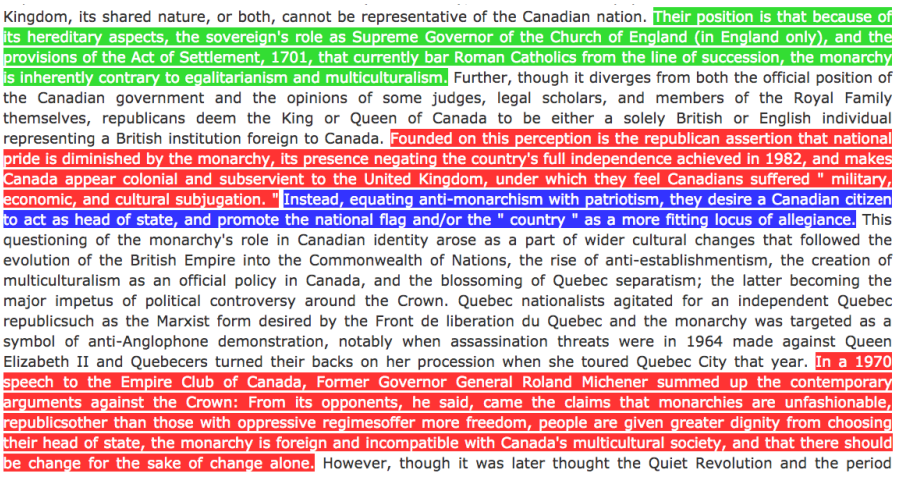
Kingdom, its shared nature, or both, cannot be representative of the Canadian nation. Their position is that because of its hereditary aspects, the sovereign's role as Supreme Governor of the Church of England (in England only), and the provisions of the Act of Settlement, 1701, that currently bar Roman Catholics from the line of succession, the monarchy is inherently contrary to egalitarianism and multiculturalism. Further, though it diverges from both the official position of the Canadian government and the opinions of some judges, legal scholars, and members of the Royal Family themselves, republicans deem the King or Queen of Canada to be either a solely British or English individual representing a British institution foreign to Canada. Founded on this perception is the republican assertion that national pride is diminished by the monarchy, its presence negating the country's full independence achieved in 1982, and makes Canada appear colonial and subservient to the United Kingdom, under which they feel Canadians suffered " military, economic, and cultural subjugation. " Instead, equating anti-monarchism with patriotism, they desire a Canadian citizen to act as head of state, and promote the national flag and/or the " country " as a more fitting locus of allegiance. This questioning of the monarchy's role in Canadian identity arose as a part of wider cultural changes that followed the evolution of the British Empire into the Commonwealth of Nations, the rise of anti-establishmentism, the creation of multiculturalism as an official policy in Canada, and the blossoming of Quebec separatism; the latter becoming the major impetus of political controversy around the Crown. Quebec nationalists agitated for an independent Quebec republicsuch as the Marxist form desired by the Front de liberation du Quebec and the monarchy was targeted as a symbol of anti-Anglophone demonstration, notably when assassination threats were in 1964 made against Queen Elizabeth II and Quebecers turned their backs on her procession when she toured Quebec City that year. In a 1970 speech to the Empire Club of Canada, Former Governor General Roland Michener summed up the contemporary arguments against the Crown: From its opponents, he said, came the claims that monarchies are unfashionable, republicsother than those with oppressive regimesoffer more freedom, people are given greater dignity from choosing their head of state, the monarchy is foreign and incompatible with Canada's multicultural society, and that there should be change for the sake of change alone. However, though it was later thought the Quiet Revolution and the period

Figure 10. Example of data automatically annotated using the framework of Lippi and Torroni on the Wikipedia page "Republicanism_in_Canada".

2.7 Evaluation of the automatic annotation step

The last step in the argument mining pipeline is evaluation of the automatic annotation. This is an analogue of the step of evaluating manual annotation: in the previous, linguistic part of the process we compared analyses of the same set of texts tagged by different annotators, and here – in a similar way – we compare an analysis created by a human annotator and stored in our test corpus with the automatic annotation created by the system. A key difference between these two types of comparison (manual vs. manual and manual vs. automatic) is that it is assumed that neither of the human annotators is better than the other (i.e. if they create two different argument maps, we do not assume that one of them is correct), whereas in the second case we treat the manual analysis as a gold standard; i.e. whenever there is a difference between human annotation and machine annotation, we assume that the manual analysis is the correct one.

A simple measure often used for this task is accuracy, i.e. the proportion (percentage) of matches between manual and machine assignment of labels. If we want to capture more detailed information about how well the system has performed in mining arguments, a group of metrics – recall, precision and F_1 score – can be used. These are defined using the following scores: true positives, tp, when the machine has assigned a label to the same text span as the human analyst; true negatives, tn, when neither the machine nor the human analyst has assigned a label to a given segment; false positives, fp, when the machine has assigned a label to a given text span while the human being did not; and false negatives, fn, when the machine did not assign a label to a segment to

```
< utterance speaker = "lj" illoc = " standard_assertion " > < textunit nb = " 215 "
> it was a ghastly aberration < /textunit> < /utterance>

<utterance speaker = ”cl” illoc = ”RQ”> <textunit nb= ”216”> or was it in fact
typical ? < /textunit> </utterance> .

<utterance speaker = ”cl” illoc = ”RQ-AQ”> <textunit nb = ”217”>
was it the product of a policy that was unsustainable that could
only be pursued by increasing repression? < /textunit> < /utterance>.
```

Figure 11. Example of data automatically annotated using the <TextCoop> platform for discourse processing of dialogical arguments in natural-language transcribed texts of the BBC Radio 4 programme *Moral Maze*

which the human being made an assignment. Then:
– **recall** measures how many times the system failed to recognise ("missed out") arguments:

$$R = \frac{tp}{tp + fn}$$

– **precision** shows how many times the program incorrectly identified a text span as an argument:

$$P = \frac{tp}{tp + fp}$$

– $\mathbf{F}_1$ **score** (F-score, F-measure) provides the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

If the matrices are computed and the performance of the system proves to be low, then we need to repeat the computational part of the process of argument mining, attempting to improve the NLP techniques and methods we are using.

For example, in their work [Moens *et al.*, 2007; Palau and Moens, 2009] Palau and Moens obtain the following F_1-scores: 0.68 for the classification of premises; 0.74 for the classification of conclusions; and 0.60 for the determination of argument structures. For the eRulemaking corpus [Park and Cardie, 2014a], the system developed by Park and Cardie has the performance $P = 86.86$, $R = 83.05$, $\mathrm{F}_1 = 84.91$ for the label Unverifiable Proposition (UnVerif) vs All; and $P = 49.88$, $R = 55.14$ and $\mathrm{F}_1 = 52.37$ for the label Verifiable Proposition Non-experiential (VERIFnon) vs All. Finally, in the work [Lawrence and Reed, 2016], Lawrence and Reed aim to use argumentation schemes and combine different techniques in order to improve the success of recognising

argument structure, achieving the following results: for the technique of Discourse Indicators the system delivers precision of 1.00, recall of 0.08, and an F_1-score of 0.15; for the technique of Topic Similarity the system has precision of 0.70, recall of 0.54 and an F_1-score of 0.61; for the technique of Schematic Structure the system delivers precision of 0.82, recall of 0.69, and an F_1-score of 0.75; and finally, for the combination of these techniques the system improves performance and delivers precision of 0.91, recall of 0.77, and an F_1-score of 0.83.

3 An example: predicting argument relations

As we discussed in the introduction of this chapter, there are many approaches to argument mining, from the perspectives of computational models of arguments or of computational linguistics. In this section, we highlight the main features of an approach to argument mining where the selected argumentation model is the abstract argumentation framework [Dung, 1995; Cayrol and Lagasquie-Schiex, 2013] and the NLP method applied is Textual Entailment.

More precisely, one of the main goals in the argument mining pipeline is to predict the relation holding between pairs of arguments. In this section, we describe in detail one of the few approaches proposed in the argument mining community to address this task [Cabrio and Villata, 2012b; Cabrio and Villata, 2013], while also showing the execution and implementation of the argument pipeline presented in Section 2. The idea is to predict positive and negative relations between arguments, i.e. *support* and *attack* relations, using textual entailment. Alongside formal approaches to semantic inference that rely on logical representation of meaning, the notion of Textual Entailment (TE) has been proposed as an applied framework to represent major semantic inferences across applications in the field of computational linguistics [Dagan *et al.*, 2009]. The development of the Web has led to a paradigm shift, due to the need to process a huge amount of available (but often noisy) data. TE is a generic framework for applied semantics, where linguistic objects are mapped by means of semantic inferences at a textual level.

3.1 The NoDE dataset

The NoDE dataset (Natural language arguments in online DEbates) is a benchmark of natural language arguments extracted from different kinds of textual sources[34]. It is composed of three datasets of natural language arguments, released in two machine-readable formats: the standard XML format and the XML/RDF format adopting the SIOC-Argumentation ontology, which has been extended in [Cabrio *et al.*, 2013] to deal with bipolar abstract argumentation.

We have identified three different scenarios for data extraction: *(i)* the online debate platforms Debatepedia[35] and ProCon[36] present a set of topics to be

[34] Available at: `www.inria.fr/NoDE`
[35] `http://idebate.org/debatabase`
[36] `http://www.procon.org/`

discussed, and participants argue about issues related to such topics that are raised on the platform, pointing out whether their arguments are in favour of or against a particular position on a given issue; *(ii)* in the script of the famous play *Twelve Angry Men*, the jurors of a trial discuss whether or not a boy is guilty, and at the end of each act they vote to determine whether they all agree; and *(iii)* the Wikipedia revision history of an article shows its evolution over time (in our case, a four-year period); we focused on the five most revised articles. These three scenarios lead to three different resources: the online debates resource collects arguments about a debated issue or other arguments into small bipolar argumentation graphs. The same happens for the Wikipedia dataset, with the revisions of the articles used to build small bipolar argumentation graphs. The "Twelve Angry Men" resource also collects pro and con arguments, but the resulting three bipolar argumentation graphs present a higher complexity than the debate graphs.

Methodology. Given a set of linked arguments (e.g. in a debate), we proceed as follows:

1. We couple each argument with the argument it is related to (which it attacks or supports). The first layer of the dataset is therefore composed of pairs of arguments (each one labelled with a unique ID) annotated by the semantic relations linking them (i.e. *attack* or *support*);

2. starting from the pairs of arguments in the first layer, we then build a bipolar argumentation graph for each topic in the dataset. In the second layer of the dataset, we therefore find argument graphs.

To create the dataset of argument pairs, we followed the criteria defined and used by the organisers of the Recognizing Textual Entailment challenge [Dagan *et al.*, 2009][37]. To test the progress of Textual Entailment (TE) systems in a comparable setting, RTE participants are provided with datasets composed of pairs of textual snippets (the Text T and the Hypothesis H) involving various levels of entailment reasoning (e.g. lexical or syntactic). The TE systems are required to produce a correct judgement on the given pairs (i.e. to say if the meaning of H can be inferred from the meaning of T). Two kinds of judgements are allowed: two-way (yes-or-no entailment) or three-way judgement (entailment, contradiction or unknown). To perform the latter, if there is no entailment between T and H, systems must be able to distinguish whether the truth of H is contradicted by T or remains unknown on the basis of the information contained in T. To correctly judge each single pair inside the RTE datasets, systems are expected to cope with both the different linguistic phenomena involved in TE and the complex ways in which they interact.

[37] Since 2004, RTE Challenges have promoted research in RTE `http://www.nist.gov/tac/2010/RTE/`.

Data format. As regards the choice of format, the NoDE dataset also uses the one proposed by the RTE challenges to annotate the first layer of the data in NoDE. It is an XML format, where each pair is identified by a unique ID and by the task (in this case, *argumentation*). The element *entailment* contains the annotated relation of entailment/non-entailment between the two arguments in the pair. Unlike the RTE dataset, in NoDE the element *topic* is added to identify the graph name to which the pair belongs, as well as an ID to keep track of each text snippet (i.e. each argument). The argument IDs are unique within each graph. An example pair from the Debatepedia dataset is as follows:

```
<entailment-corpus>
  <pair task="ARG" id="1" topic="Violentgames" entailment=
  "NONENTAILMENT">
      <t id="2">Violent video games do not increase aggression.</t>
      <h id="1">Violent games make youth more aggressive/violent.
      </h>
  </pair>
    ...
</entailment-corpus>
```

For the second layer, the XML/RDF format adopted in NoDE relies on the SIOC-Argumentation extended vocabulary[38]. Each argument is a `sioc_arg:Argument`, and the two relations of *support* and *attack* are respectively characterised by the properties `sioc_arg:supportsArg` and `sioc_arg:challengesArg` (mapped on the *entailment* and *nonentailment* relations, respectively). An example from the Twelve Angry Men XML/RDF dataset is as follows:

```
<http://example.org/12AngryMen/pair1t> rdf:type sioc_arg:Argument ;
  sioc:content "Ever since he was five years old his father beat
                him up regularly. He used his fists." ;
  sioc_arg:challengesArg <http://example.org/12AngryMen/pair1h> .

<http://example.org/12AngryMen/pair1h> rdf:type sioc_arg:Argument ;
  sioc:content "Look at the kid's record. At fifteen he was
  in reform  school. He stole a car. He was picked up for
  knife-fighting. I think they said he stabbed somebody in the arm.
  This is a very fine boy." .
```

All the abstract bipolar argumentation graphs resulting from the datasets of the benchmark are also available for visualisation as `png` images.

Debatepedia/ProCon dataset. To build the first benchmark of natural language arguments, the Debatepedia and ProCon encyclopaedia of pro and con arguments on critical issues were selected as data sources. To fill in the

[38] `http://bit.ly/SIOC_Argumentation`

first layer of the dataset, a set of topics (Table 5 column *Topics*) of Debatepedia/ProCon debates was selected, and for each topic, the following procedure was applied:

1. The main issue (i.e. the title of the debate expressed as an affirmative statement) is considered as the starting argument.

2. Each user opinion is extracted and considered an argument.

3. Since *attack* and *support* are binary relations, the arguments are coupled with either

   (a) the starting argument, or

   (b) other arguments in the same discussion to which the most recent argument refers (i.e., when a user's opinion supports or attacks an argument previously expressed by another user, the former is coupled with the latter), in chronological order to maintain the dialogue structure.

4. The resulting pairs of arguments are then tagged with the appropriate relation, i.e. *attack* or *support*.

The use of Debatepedia/ProCon as a case study provides us with arguments that have already been annotated (*pro* ⇒ *entailment*, and *con* ⇒ *contradiction*) and casts our task as a yes/no entailment task. To show a step-by-step application of the procedure, let us consider the debated issue *Can coca be classified as a narcotic?* In step 1, we transform its title into an affirmative form and consider it to be the starting argument (a). Then, in step 2, we extract all the users' opinions concerning this issue (both pro and con), e.g. (b), (c) and (d):

(4) a. *Coca can be classified as a narcotic.*
   b. *In 1992 the World Health Organization's Expert Committee on Drug Dependence (ECDD) undertook a "prereview" of coca leaf. The ECDD report concluded that, "the coca leaf is appropriately scheduled as a narcotic under the Single Convention on Narcotic Drugs, 1961, since cocaine is readily extractable from the leaf." This ease of extraction makes coca and cocaine inextricably linked. Therefore, because cocaine is defined as a narcotic, coca must also be defined in this way.*
   c. *Coca in its natural state is not a narcotic. What is absurd about the 1961 convention is that it considers the coca leaf in its natural, unaltered state to be a narcotic. The paste or the concentrate that is extracted from the coca leaf, commonly known as cocaine, is indeed a narcotic, but the plant itself is not.*

d. *Coca is not cocaine. Coca is distinct from cocaine. Coca is a natural leaf with very mild effects when chewed. Cocaine is a highly processed and concentrated drug using derivatives from coca, and therefore should not be considered as a narcotic.*

In step 3a we couple arguments (b) and (d) with the starting issue, since they are directly linked to it, and in step 3b we couple argument (c) with argument (b) and argument (d) with argument (c), since they follow one another in the discussion (i.e. the user expressing argument (c) answers the user expressing argument (b), so that the arguments are concatenated; the same applies to arguments (d) and (c)).

In step 4, the resulting pairs of arguments are then tagged with the appropriate relation: **(b)** *supports* **(a)**, **(d)** *attacks* **(a)**, **(c)** *attacks* **(b)** and **(d)** *supports* **(c)**.

We collected 260 T-H pairs (Tables 1, 4). The training set is composed of 85 entailment and 75 contradiction pairs and the test set consists of 55 entailment and 45 contradiction pairs. Pairs in the test set are extracted from topics different from those of the training set.

Training set					**Test set**				
Topic	**#arg**	**#pairs**			**Topic**	**#arg**	**#pairs**		
		tot	**yes**	**no**			**tot**	**yes**	**no**
Violent games	16	15	8	7	*Ground zero mosque*	9	8	3	5
China 1-child policy	11	10	6	4	*Military service*	11	10	3	7
Coca as a narcotic	15	14	7	7	*Libya no fly zone*	11	10	6	4
Child beauty contests	12	11	7	4	*Airport security prof.*	9	8	4	4
Arming Libyan rebels	10	9	4	5	*Solar energy*	16	15	11	4
Alcohol breath tests	8	7	4	3	*Natural gas vehicles*	12	11	5	6
Osama death photo	11	10	5	5	*Cell phones/driving*	11	10	5	5
Private social security	11	10	5	5	*Legalize marijuana*	17	16	10	6
Internet as a right	15	14	9	5	*Gay marriages*	7	6	4	2
Tablets vs. Textbooks	22	21	11	10	*Vegetarianism*	7	6	4	2
Obesity	16	15	7	8					
Abortion	25	24	12	12					
TOTAL	172	**160**	**85**	**75**	**TOTAL**	110	**100**	**55**	**45**

Table 3. The Debatepedia dataset

Based on the TE definition [Dagan *et al.*, 2009], an annotator with skills in linguistics carried out the first phase of annotation of the Debatepedia/ProCon dataset. Then, to assess the validity and reliability of the annotation, the same annotation task was independently carried out by a second annotator, so that inter-annotator agreement could be computed. As discussed in Section 2.4, Cohen's kappa was used to calculate inter-rater agreement on the NoDE dataset. Inter-rater agreement results in $\kappa = 0.7$ (see Table 4). As a rule of thumb, this is a satisfactory level of agreement, so we consider these annotated datasets as the *gold standard*, i.e. a reference dataset to which the performances of automated systems can be compared.

To fill the second layer of the Debatepedia/ProCon dataset, the pairs annotated in the first layer are combined to build a bipolar argumentation graph for each topic (12 topics in the training set and 10 topics in the test set, listed in Table 5).

Twelve Angry Men dataset. As a second scenario for extraction of natural language arguments, the NoDE dataset selected the script of the play *Twelve Angry Men*. The play concerns the deliberations of a jury in a homicide trial. The story begins after closing arguments have been presented in the homicide case. At first, the jurors have a nearly unanimous decision that the defendant is guilty, with a single dissenter voting "not guilty" and sowing a seed of reasonable doubt throughout the play.

The play is divided into three acts, and at the end of each act the jury votes, until they reach unanimity. For each act, we manually identified the arguments (excluding sentences which cannot be considered self-contained arguments) and coupled each argument with the argument it supports or attacks in the dialogue flow (Examples (5) and (6), respectively). In the discussions, one character's argument comes after another's (entailing or contradicting one of the arguments previously expressed by another character); therefore, pairs are created in the graph connecting the former to the latter (more recent arguments are considered to be T and the argument with respect to which we want to detect the relation is considered to be H). In Example (7), juror 1 claims argument (7-c) and is attacked by juror 2, claiming argument (7-b). Juror 3 then claims argument (7-a) to support juror 2's opinion. In the dataset, the following couples were annotated: (7-c) is contradicted by (7-b) and (7-b) is entailed by (7-a). In Example (8), juror 1 claims argument (8-c) supported by juror 2 (argument (8-b)) and juror 3 attacks juror's 2 opinion with argument (8-a). More specifically, (8-c) is entailed by (8-b) and (8-b) is contradicted by (8-a).

(5) a. *Maybe the old man didn't hear the boy yelling "I'm going to kill you". I mean with the el noise.*
b. *I don't think the old man could have heard the boy yelling.*

(6) a. *I never saw a guiltier man in my life. You sat right in court and heard the same thing I did. The man's a dangerous killer.*
b. *I don't know if he is guilty.*

(7) a. *Maybe the old man didn't hear the boy yelling "I'm going to kill you". I mean with the el noise.*
b. *I don't think the old man could have heard the boy yelling.*
c. *The old man said the boy yelled "I'm going to kill you" out. That's enough for me.*

(8) a. *The old man cannot be a liar, he must have heard the boy yelling.*
b. *Maybe the old man didn't hear the boy yelling "I'm going to kill you". I mean with the el noise.*
c. *I don't think the old man could have heard the boy yelling.*

Given the complexity of the play and the fact that in human linguistic interactions much is left implicit, we simplified the arguments as follows: *i)* by adding the required context in T to make the pairs self-contained (in the TE framework entailment is detected based on the evidence provided in T); and *ii)* by solving intra-document co-references, as in *Nobody has to prove that!*, transformed into *Nobody has to prove [that he is not guilty].*
A total of 80 T-H pairs were collected: 25 entailment pairs, 41 contradiction pairs and 14 unknown pairs (contradiction and unknown pairs can be collapsed in the judgement *non-entailment* for the two-way classification task). Table 4 shows the inter-annotator agreement.

To fill the second layer of the Twelve Angry Men dataset, again the pairs annotated in the first layer are combined to build a bipolar argumentation graph for each topic in the dataset (the three acts of the play). The complexity of the graphs obtained for this scenario is higher than in the debate graphs (on average, 27 links per graph as compared to 9 links per graph in the Debatepedia dataset).

Wikipedia revision history dataset. We selected four dumps of English Wikipedia (*Wiki 09* dated 6.03.2009, *Wiki 10* dated 12.03.2010, *Wiki 11* dated 9.07.2011, and *Wiki 12* dated 6.12.2012), and NoDE focuses on the five most revised pages[39] at that time (George W. Bush, United States, Michael Jackson, Britney Spears, and World War II).

After extracting plain text from these pages, each document was sentence-split for both *Wiki 09* and *Wiki 10*, and the sentences of the two versions were automatically aligned to create pairs. Then, following [Cabrio *et al.*, 2012], the *Position Independent Word Error Rate (PER)* – a metric based on calculation of the number of words which differ between a pair of sentences – was adopted in order to measure the similarity between the sentences in each pair. Only pairs composed of sentences where major editing was carried out ($0.2 < PER < 0.6$), but still describing the same event, were selected. For each pair of extracted sentences, TE pairs were created by setting the revised sentence (from *Wiki 10*) as T and the original sentence (from *Wiki 09*) as H. Starting from such pairs composed of the same revised argument, we checked the more recent Wikipedia versions (i.e. *Wiki 11* and *Wiki 12*) to see whether the arguments had been further modified. If so, another T-H pair based on the same assumptions as before was created by setting the revised sentence as T and the older sentence as H (see Example (9)).

(9) a. Wiki12: *The land area of the contiguous United States is 2,959,064 square miles (7,663,941 km2).*
b. Wiki11: *The land area of the contiguous United States is approximately 1,900 million acres (7,700,000 km2).*
c. Wiki10: *The land area of the contiguous United States is approximately 1.9 billion acres (770 million hectares).*

[39] `http://bit.ly/WikipediaMostRevisedPages`

d. Wiki09: *The total land area of the contiguous United States is approximately 1.9 billion acres.*

Such pairs were then annotated with respect to the TE relation, following the criteria defined for the two-way judgement TE task. As a result of the first step (extraction of the revised arguments in (9-d) and (9-c)), 280 T-H pairs were collected. After applying the procedure to the same arguments in (9-b) and (9-a), the total number of collected pairs was 452 (training set composed of 114 entailment and 114 non-entailment pairs and test set of 101 entailment and 123 non-entailment pairs); see Table 4. To enable correct training of automatic systems, the dataset was balanced with respect to the percentage of yes/no judgements. In Wikipedia, the actual distribution of attacks and supports among revisions of the same sentence is slightly unbalanced, as users generally edit a sentence to add different information or correct it. Inter-annotator agreement is reported in Table 4.

Dataset	**pairs**	**supp**	**att**	**graphs**	**Inter-annotator agreement**		
					annotators	**pairs**	κ
Debatepedia	260	140	120	22	2	100	0.7
12 Angry Men	80	25	55	3	2	40	0.74
Wiki revisions	452	215	237	416	2	140	0.7

Table 4. Summary of the datasets composing NoDE, where *supp* means support and *att* means attack

3.2 Textual Entailment for relation prediction in abstract argumentation

In the reminder of this section, we will provide some basic information about Textual Entailment and abstract bipolar argumentation theory, and then we will detail how relation prediction in the argument mining pipeline can be addressed using TE systems. An evaluation of the results obtained by such a method are reported as well.

Textual Entailment: background and main insights. Classical approaches to semantic inference rely on logical representations of meaning that are external to the language itself and are typically independent of the structure of any particular natural language. Texts are first translated or interpreted into some logical form and then new propositions are inferred from interpreted texts by a logic theorem prover. But, especially since the development of the Web, we have witnessed a paradigm shift, due to the need to process a huge amount of available (but often noisy) data. Addressing the inference task by means of logic theorem provers in automated applications aimed at natural language understanding has shown several intrinsic limitations [Blackburn *et al.*, 2001].

Especially in data-driven approaches, such as the one presented in this work, where patterns are learnt from large-scale naturally-occurring data, we can accept approximate answers provided by efficient and robust systems, even at the price of logical unsoundness or incompleteness. Starting from these considerations, Monz and de Rijke [Monz and de Rijke, 2001] propose addressing the inference task directly at the textual level instead, exploiting currently available NLP techniques. While methods for automated deduction assume that the arguments in the input are already expressed in some formal representation of meaning (e.g. first-order logic), addressing the inference task at a textual level opens new and different challenges from those encountered in formal deduction. More emphasis is placed on informal reasoning, lexical semantic knowledge, and the variability of linguistic expressions (see Section 2).

The notion of Textual Entailment has been proposed as an applied framework to detect major semantic inferences across applications in NLP [Dagan *et al.*, 2009]. It is defined as a relation between a coherent textual fragment (the Text T) and a language expression, which is considered the Hypothesis (H). Entailment holds (i.e. $T \Rightarrow H$) if the meaning of H can be inferred from the meaning of T, as interpreted by a typical language user. The TE relationship is directional, since the meaning of one expression usually entails the other, while the opposite is much less certain. Consider the pairs in Examples (10) and (11).

(10) a. **T1** *Internet access is essential now; must be a right. The internet is only that wire that delivers freedom of speech, freedom of assembly, and freedom of the press in a single connection.*
b. **H** *Making Internet a right only benefits society.*

(11) a. **T2** *Internet not as important as real rights. We may think of such trivial things as a fundamental right, but consider the truly impoverished and what is most important to them. The right to vote, the right to liberty and freedom from slavery or the right to elementary education.*
b. **H** *Making Internet a right only benefits society.*

A system aimed at recognising TE should detect an inference relation between T1 and H (i.e. the meaning of H can be derived from the meaning of T) in Example (10), while it should not detect an entailment between T2 and H in Example (11). The definition of TE is based on (and assumes) common human understanding of language, as well as common background knowledge. However, the entailment relation is said to hold only if the statement in the text licenses the statement in the hypothesis, meaning that the content of T and common knowledge together should entail H, and not background knowledge alone. In this applied framework, inferences are performed directly over lexical-syntactic representations of the texts.

For pairs where the entailment relation does not hold between T and H, systems are required to make a further distinction between pairs where the en-

tailment does not hold because the content of H is contradicted by the content of T (*contradiction*, see Example (11)), and pairs where the entailment cannot be determined because the truth of H cannot be verified on the basis of the content of T (*unknown*, see Example (12)). De Marneffe and colleagues [de Marneffe *et al.*, 2008] provide a definition of contradiction for the TE task, claiming that it occurs when two sentences *i)* are extremely unlikely to be true simultaneously and *ii)* involve the same event. This three-way judgment task (*entailment* vs *contradiction* vs *unknown*) was introduced for the RTE-4 challenge. Before RTE-4, TE was considered a two-way decision task (*entailment* vs *no entailment*). However, the classic two-way task is also offered as an alternative in recent editions of the evaluation campaign (*contradiction* and *unknown* judgments are collapsed into the judgment *no entailment*).

(12) a. **T3** *Internet "right" means denying parents' ability to set limits. Do you want to make a world when a mother tells her child: "you cannot stay on the internet anymore" that she has taken a right from him? Compare taking the right for a home or for education with taking the "right" to access the internet.*

b. **H** *Internet access is essential now; must be a right. The internet is only that wire that delivers freedom of speech, freedom of assembly, and freedom of the press in a single connection.*

Abstract bipolar argumentation. Abstract argumentation frameworks [Dung, 1995] consider arguments as abstract entities, deprived of any structural property and of any relations but attack. However, in these frameworks, the assessment of argument acceptability depends only on the attack relation in abstract terms, while in other application domains other relations may be required. In particular, abstract bipolar argumentation frameworks, first proposed by [Cayrol and Lagasquie-Schiex, 2005], extend Dung's abstract framework by taking into account both the attack relation and the support relation. An abstract bipolar argumentation framework is a labelled directed graph, with two labels, indicating either attack or support. We represent the attack relation by $a \rightarrow b$ and the support relation by $a \dashrightarrow b$.

Casting bipolar argumentation as a TE problem. The issue of predicting the relation holding between two arguments can be instantiated in the task of predicting whether the relation between two arguments is a support or attack relation. In the remainder of this section, we present and evaluate this approach on the data of the NoDE dataset. For more details about this task and its main insights, we refer the reader to [Cabrio and Villata, 2012a; Cabrio and Villata, 2013; Cabrio *et al.*, 2013; Cabrio and Villata, 2014].

The general goal of our work is to propose an approach to help the participants in forums or debates (e.g. on Debatepedia or Twitter) detect which arguments about a certain topic are accepted. As a first step, we need to *(i)* automatically generate the arguments (i.e., recognise a participant's opinion on a certain topic as an argument) and *(ii)* detect their relation to the other

arguments. We cast the problem as a TE problem, where the T-H pair is a pair of arguments expressed by two different participants in a debate on a certain topic. For instance, given the argument "Making Internet a right only benefits society" (which we regard as H), participants can be in favour of it (expressing arguments from which H can be inferred, as in Example (10)), or can contradict it (expressing an opinion against it, as in Example (11)). Since in debates one participant's argument comes after another's, we can extract such arguments and compare them both w.r.t. the main issue and w.r.t. the other participants' arguments (when the new argument entails or contradicts one of the arguments previously expressed by another participant). For instance, given the same debate as before, a new argument T3 may be expressed by a third participant to contradict T2 (which becomes the new H (H1) in the pair), as shown in Example (13).

(13) a. **T3** *I've seen the growing awareness within the developing world that computers and connectivity matter and can be useful. It's not that computers matter more than water, food, shelter and healthcare, but that the network and PCs can be used to ensure that those other things are available. Satellite imagery sent to a local computer can help villages find fresh water, mobile phones can tell farmers the prices at market so they know when to harvest.*

b. **T2 ≡ H1** *Internet not as important as real rights. We may think of such trivial things as a fundamental right, but consider the truly impoverished and what is most important to them. The right to vote, the right to liberty and freedom from slavery or the right to elementary education.*

TE provides us with techniques to identify the arguments in a debate and to detect which kind of relation underlies each pair of arguments. A TE system returns a judgement (entailment or contradiction) on the argument pairs related to a certain topic, which are used as input to build the abstract argumentation framework. Example (14) presents how TE is combined with bipolar argumentation to compute the set of accepted arguments at the end of the argument mining pipeline.

(14) *The textual entailment phase returns the following pairs for the natural language opinions detailed in Examples (10), (11), and (13):*

- *T1 entails H*
- *T2 attacks H*
- *T3 attacks H1 (i.e. T2)*

Given this result, the argumentation module of our framework maps each element to its corresponding argument: H ≡ A_1, T1 ≡ A_2, T2 ≡ A_3, and T3 ≡ A_4. The accepted arguments (using admissibility-based semantics) are $\{A_1, A_2, A_4\}$. This means that the issue "Making Internet a right only benefits society"' A_1 is considered to be accepted.

Experimental evaluation: the Debatepedia/ProCon dataset of NoDE. Following the methodology described in Section 2.7, in the first set of experiments the 200 T-H pairs of the Debatepedia/ProCon dataset have been divided into 100 for training and 100 for testing of the TE system (each dataset is composed of 55 entailment and 45 contradiction pairs). The pairs considered for the test set concern completely new topics, never seen by the system. Table 5 shows the topics used to train the system and those used to test it.

Training set				
Topic	**#argum**	**#pairs**		
		TOT.	**yes**	**no**
Violent games boost aggressiveness	16	15	8	7
China one-child policy	11	10	6	4
Consider coca as a narcotic	15	14	7	7
Child beauty contests	12	11	7	4
Arming Libyan rebels	10	9	4	5
Random alcohol breath tests	8	7	4	3
Osama death photo	11	10	5	5
Privatizing social security	11	10	5	5
Internet access as a right	15	14	9	5
TOTAL	109	**100**	**55**	**45**
Test set				
Topic	**#argum**	**#pairs**		
		TOT.	**yes**	**no**
Ground zero mosque	9	8	3	5
Mandatory military service	11	10	3	7
No fly zone over Libya	11	10	6	4
Airport security profiling	9	8	4	4
Solar energy	16	15	11	4
Natural gas vehicles	12	11	5	6
Use of cell phones while driving	11	10	5	5
Marijuana legalization	17	16	10	6
Gay marriage as a right	7	6	4	2
Vegetarianism	7	6	4	2
TOTAL	110	**100**	**55**	**45**

Table 5. The Debatepedia dataset used in our experiments

To detect which kind of relation underlies each pair of arguments, we adopted the modular architecture of the EDITS system (Edit Distance Textual Entailment Suite), version 3.0. EDITS is an open-source software package for recognising TE[40] [Kouylekov and Negri, 2010] which implements a distance-based

[40] `http://edits.fbk.eu/`

		Train			**Test**		
	rel	*Pr.*	*Rec.*	*Acc.*	*Pr.*	*Rec.*	*Acc.*
EDITS	*yes*	0.71	0.73	**0.69**	0.69	0.72	**0.67**
	no	0.66	0.64		0.64	0.6	
WordOverl.	*yes*	0.64	0.65	0.61	0.64	0.67	0.62
	no	0.56	0.55		0.58	0.55	

Table 6. System performance on the Debatepedia/ProCon dataset of NoDE (precision, recall and accuracy)

framework assuming that the probability of an entailment relation between a given T-H pair is inversely proportional to the distance between T and H (i.e., the higher the distance, the lower the probability of entailment)[41].

A two-step evaluation is then carried out. First, we need to assess the performance of the TE system in assigning entailment and contradiction relations to the pairs of arguments in the Debatepedia/ProCon dataset. Then, we evaluate to what degree its performance affects the application of the argumentation theory module, i.e. to what degree an incorrect assignment of a relation to a pair of arguments is propagated in the argumentation framework.

For the first evaluation, we run EDITS on the Debatepedia/ProCon training set so it can learn the model, and then test it on the test set. EDITS was tuned in the following configuration: *i)* cosine similarity as the core distance algorithm, *ii)* distance calculated on lemmas, and *iii)* a stopword list defined to set no distance between stopwords. We used the system off-the-shelf, applying one of its basic configurations. Table 6 reports on the results obtained both using EDITS and using a baseline that applies a Word Overlap algorithm on tokenised text. Even where a basic configuration of EDITS and a small data set (100 pairs for training) were used, performance on the Debatepedia/ProCon test set is promising and in line with those of TE systems on RTE data sets (usually containing about 1,000 pairs for training and 1,000 for testing).

As a second step in the evaluation phase, we need to consider the impact of EDITS performances on the acceptability of the arguments, i.e. to what extent incorrect assignment of a relation to a pair of arguments affects the acceptability of the arguments in the argumentation framework. Admissibility-based semantics are then used to identify the accepted arguments for the argumentation frameworks of the gold standard and those resulting from the TE system. The precision of the combined approach to identification of accepted arguments (i.e. arguments accepted by the combined system and by the gold standard with regard to a given Debatepedia/ProCon topic) is on average 0.74, and the

[41] In previous RTE challenges, EDITS has always ranked among the 5 best participating systems out of an average of 25 systems, and is one of the few RTE systems available as open source `http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool`.

Table 7. System performances on the Wikipedia revision history dataset of NoDE

		Train			**Test**		
EDITS conf.	*rel*	*Pr.*	*Rec.*	*Acc.*	*Pr.*	*Rec.*	*Acc.*
WordOverlap	*yes*	0.83	0.82	**0.83**	0.83	0.82	**0.78**
	no	0.76	0.73		0.79	0.82	
CosineSimilarity	*yes*	0.58	0.89	0.63	0.52	0.87	0.58
	no	0.77	0.37		0.76	0.34	

recall (arguments accepted in the gold standard and retrieved as accepted by the combined system) is 0.76. Its accuracy (the ability of the combined system to accept some arguments and discard others) is 0.75, meaning that the TE system's mistakes in assigning relations propagate in the argumentation framework, but the results are still satisfactory.

Experimental evaluation: the Wikipedia revision history dataset of NoDE. To evaluate the results of relation prediction by means of Textual Entailment with respect to the Wikipedia revision history dataset of NoDE, we again need to carry out a two-step evaluation: first, we must assess the performance of EDITS in assigning the *entailment* and the *no entailment* relations to pairs of arguments with respect to the Wikipedia dataset, and then we can evaluate the extent to which the performance affects the application of the argumentation theory module, i.e. the extent to which incorrect assignment of a relation to a pair of arguments is propagated in the argumentation framework. For the first evaluation, EDITS was run on the Wikipedia training set to learn the model, and then tested on the test set. In the configuration of EDITS, the distance entailment engine applies *cosine similarity* and *word overlap* as the core distance algorithms. In both cases, distance is calculated on lemmas and a stopword list is defined to have no distance value between stopwords.

The results are reported in Table 7. Due to the specificity of this dataset (i.e., it is composed of revisions of arguments), the *word overlap* algorithm outperforms *cosine similarity*, as there is high similarity between the revised and original arguments (in most of the positive examples the two sentences are very close or there is an almost perfect inclusion of H in T). For the same reason, the results are higher than, for example, in Cabrio and Villata [Cabrio and Villata, 2012a], and higher than the results obtained on average in RTE challenges. For these runs, we used the system off-the-shelf, applying its basic configuration. As a second step in the evaluation phase, we consider the impact of EDITS performances (obtained using word overlap, since it provided the best results) on the acceptability of the arguments, i.e. the extent to which incorrect assignment of a relation to a pair of arguments affects the acceptability of the arguments in the argumentation framework. We again use admissibility-based

semantics [Dung, 1995] to identify the accepted arguments both with respect to the correct argumentation frameworks of each Wikipedia revised argument (where entailment/contradiction relations are correctly assigned, i.e. the gold standard) and on the frameworks generated by assigning the relations resulting from the TE system judgements. The precision of the combined approach in identifying accepted arguments (i.e. arguments accepted by the combined system and by the gold standard with respect to a specific Wikipedia revised argument) is on average 0.90 and the recall (arguments accepted in the gold standard and retrieved as accepted by the combined system) is 0.92. The F-measure is 0.91, meaning that the TE system's mistakes in relation assignment propagate in the argumentation framework, but results are still satisfactory.

4 Conclusions

In this chapter, we have presented the entire argument mining pipeline with an in-depth look at two approaches proposed in the literature, applied to structured and abstract argumentation (supporting arguments and conflicting arguments). A case study of automatic extraction of argument relations is then discussed in Section 3.

This chapter has highlighted rising trends in the very new field of argument mining research. We can summarise the main steps of the argument mining pipeline as follows: *(i)* We must select a precise argument mining task and define precise guidelines for annotation of the data to be used by our system. If the guidelines are not sufficiently clear, the results of the annotation phase will be poor, leading to unsatisfactory inter-annotator agreement. Defining precise annotation guidelines is a time-consuming task, but it ensures the reliability of the resources produced when the guidelines are followed. *(ii)* When the guidelines have been established, the corpus of natural language arguments must be annotated according to these guidelines. After the annotation process, inter-annotator agreement must be evaluated to ensure the reliability of the resources. *(iii)* Now that the data is ready, we must choose or define the best solution for the task we are addressing, e.g. argument classification or relation prediction. *(iv)* Finally, the results must be evaluated to ensure that the proposed solution is correct and scalable.

Some conclusions can be drawn from this overview of the field of argument mining. First of all, it is important to distinguish between the well-known NLP research field of *opinion mining* (or *sentiment analysis*) and argument mining. Apart from some minor differences, the main point here is that the goal of opinion mining is to understand *what* people think about something, while the goal of argument mining is to understand *why* people think something about a given topic [Habernal *et al.*, 2014]. This key difference between these two research areas characterises the main feature of argumentation theory in general, which is the ability to explain and to justify different viewpoints. Second, argumentation theory is traditionally discussed in correlation with what is called *critical thinking*, i.e. the intellectual process of objective analysis and evalu-

ation of an issue in order to form a judgement. However, argument mining approaches can support formal argumentation approaches in order to define formal models that are closer to human reasoning, in which the fuzziness and ambiguity of natural language play an important role and the intellectual process is not always completely rational and objective. Argument mining can provide greater insight into the answers to such questions as "What are the best arguments to influence a real audience?" or "What is the role of emotions in the argumentation process?"

As discussed in recently published surveys of argument mining [Peldszus and Stede, 2013; Lippi and Torroni, 2015a], argument mining approaches currently face two main issues: big data and deep learning. Concerning the first point, a huge amount of data is now available on the Web, such as social network posts, forums, blogs, product reviews, or user comments to newspapers articles, and it must be analysed automatically as it far exceeds human capabilities to parse and understand it without an automatic support tool. Argument mining can make the difference here, and can exploit the Web to perform crowd-sourced annotations for very large corpora. A first step in this direction is described in the use case in this chapter, based on the NoDE dataset, whose texts are extracted from various online sources. As shown in this example, classical semantic approaches to NLP cannot cope with the variability and noise present in such texts and the huge quantity of data that must be processed. Hence the need to apply new methods such as Textual Entailment. Concerning the second point, deep learning methods, i.e. fast and efficient machine learning algorithms such as word embeddings[42], can be exploited in the argument mining pipeline to deal with large corpora and unsupervised learning. An issue associated with data on the Web is multilingualism, which poses a challenge to argument mining approaches. As highlighted in this chapter, the vast majority of research proposed in the field of argument mining deals with English data only. This is because far better algorithms are available for NLP tools for English texts than for other languages. However, the study of other languages may lead to the identification of language-specific argumentation patterns that would be useful for further analysis of text snippets. Other issues to be considered are how to extract an uncertainty measure from natural language arguments, e.g. those exchanged in a debate, such that this measure can be used to weight the acceptability of an argument, and determination of the most suitable visual representation of the results of the argument mining pipeline, e.g. identified argument components, argument boundaries, and relations between arguments, in order to support human decision-making processes.

Although natural language argumentation has been investigated for many centuries in philosophy and rhetoric, and has recently been studied extensively in natural language processing and computational models of argument as well, argument mining has in fact just begun to tackle this important yet challenging

[42] Automatically learned feature spaces encoding high-level, rich linguistic similarity between terms.

topic, leaving us with many interesting research questions and tasks for the future work.

Acknowledgements

Some research reported in this chapter was supported in part by the EPSRC in the UK under grant EP/M506497/1 and in part by the European Union's Horizon 2020 research and innovation programme under Marie Skodowska-Curie grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts". We would also like to thank Rory Duthie and Elena Cabrio for their useful comments and consultation.

BIBLIOGRAPHY

[Aharoni *et al.*, 2014] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[Aristotle, 2004] Aristotle. *Rhetoric*. Dover Publications, 2004.

[Ashley and Walker, 2013] Kevin D. Ashley and Vern R. Walker. Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In Enrico Francesconi and Bart Verheij, editors, *International Conference on Artificial Intelligence and Law, ICAIL '13, Rome, Italy, June 10-14, 2013*, pages 176–180. ACM, 2013.

[Austin, 1962] J. L. Austin. *How to Do Things with Words*. Oxford: Clarendon, 1962.

[Bex *et al.*, 2012] Floris Bex, Thomas Gordon, John Lawrence, and Chris Reed. Interchanging arguments between Carneades and AIF. In B. Verheij, S. Szeider, and S. Woltran, editors, *Computational Models of Argument (COMMA)*, volume 245, pages 390–397. IOS Press, 2012.

[Bex *et al.*, 2013] F. Bex, S. Modgil, H. Prakken, and Chris Reed. On logical specifications of the Argument Interchange Format. *Journal of Logic and Computation*, 23(5):951–989, 2013.

[Blackburn *et al.*, 2001] P. Blackburn, J. Bos, M. Kohlhase, and H. de Nivelle. Inference and computational semantics. *Studies in Linguistics and Philosophy, Computing Meaning*, 77(2):1128, 2001.

[Boltuzic and Snajder, 2016] Filip Boltuzic and Jan Snajder. Fill the gap! analyzing implicit premises between claims from online debates. In *Proceedings of the 3rd Workshop on Argument Mining*, 2016.

[Budon *et al.*, 2014] M. Budon, M. Lucero, and G. Simari. An aif-based labeled argumentation framework. In *Proc. of Foundations of Information and Knowledge Systems (FoIKS 2014)*, volume 8367, pages 117–135. LNCS Springer, 2014.

[Budzynska and Reed, 2011] Katarzyna Budzynska and Chris Reed. Whence inference. Technical report, University of Dundee, 2011.

[Budzynska *et al.*, 2014a] Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint Dizier, Manfred Stede, and Olena Yaskorska. Towards Argument Mining from Dialogue. In *Frontiers in Artificial Intelligence and Applications*, volume 266, pages 185–196. Computational Models of Argument (COMMA14), IOS Press, September 2014.

[Budzynska *et al.*, 2014b] Katarzyna Budzynska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. A model for processing illocutionary structures and argumentation in debates. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 917–924, 2014.

[Budzynska *et al.*, 2015] Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, Olena Yaskorska, and Barbara Konat. Automatically identifying transitions between locutions in dialogue. In *European Conference on Argumentation (ECA)*, 2015.

[Budzynska *et al.*, 2016] Katarzyna Budzynska, Mathilde Janier, Chris Reed, and Patrick Saint-Dizier. Theoretical foundations for illocutionary structure parsing. *Argument and Computation*, 7(1):91–108, 2016.

[Cabrio and Villata, 2012a] E. Cabrio and S. Villata. Natural language arguments: A combined approach. In *Procs of ECAI, Frontiers in Artificial Intelligence and Applications 242*, pages 205–210, 2012.

[Cabrio and Villata, 2012b] Elena Cabrio and Serena Villata. Generating abstract arguments: a natural language approach. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 454–461. IOS Press, 2012.

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

[Cabrio and Villata, 2014] Elena Cabrio and Serena Villata. Node: A benchmark of natural language arguments. In Simon Parsons, Nir Oren, Chris Reed, and Federico Cerutti, editors, *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*, pages 449–450. IOS Press, 2014.

[Cabrio *et al.*, 2012] E. Cabrio, B. Magnini, and A. Ivanova. Extracting context-rich entailment rules from wikipedia revision history. In *The People's Web Meets NLP Workshop*, 2012.

[Cabrio *et al.*, 2013] E. Cabrio, S. Villata, and F. Gandon. A support framework for argumentative discussions management in the web. In *Procs of The Semantic Web: Semantics and Big Data, ESWC 2013*, pages 412–426, 2013.

[Cayrol and Lagasquie-Schiex, 2005] C. Cayrol and M.C. Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *Procs of ECSQARU, LNCS 3571*, pages 378–389, 2005.

[Cayrol and Lagasquie-Schiex, 2010] C. Cayrol and M.C. Lagasquie-Schiex. Coalitions of arguments: A tool for handling bipolar argumentation frameworks. *Int. J. Intell. Syst.*, 25(1):83–109, 2010.

[Cayrol and Lagasquie-Schiex, 2013] C. Cayrol and M.-C. Lagasquie-Schiex. Bipolarity in argumentation graphs: Towards a better understanding. *Int. J. Approx. Reasoning*, 54(7):876–899, 2013.

[Chesnevar *et al.*, 2006] C. Chesnevar, J. McGinnis, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, and S. Willmott. Towards an argument interchange format. *The Knowledge Engineering Review*, 21(4):293–316, 2006.

[Cohen, 1960] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:3746, 1960.

[Dagan *et al.*, 2009] I. Dagan, B. Dolan, B. Magnini, and D. Roth. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering (JNLE)*, 15(04):i–xvii, 2009.

[de Marneffe *et al.*, 2008] Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. Finding contradictions in text. In *Procs of ACL*, 2008.

[Dung, 1995] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

[Duthie *et al.*, 2016a] R. Duthie, J. Lawrence, K Budzynska, and C. Reed. The CASS technique for evaluating the performance of argument mining. In *Proceedings of the Third Workshop on Argumentation Mining*, Berlin, 2016. Association for Computational Linguistics.

[Duthie *et al.*, 2016b] Rory Duthie, Katarzyna Budzynska, and Chris Reed. Mining ethos in political debate. In *Proceedings of 6th International Conference on Computational Models of Argument (COMMA 2016)*. IOS Press, Frontiers in Artificial Intelligence and Applications, 2016.

[Feng and Hirst, 2011] Vanessa Wei Feng and Graeme Hirst. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2011)*, pages 987–996, 2011.

[Fournier and Inkpen, 2012] Chris Fournier and Diana Inkpen. Segmentation similarity and agreement. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 152161. Association for Computational Linguistics, 2012.

[Freeman, 1991] James B Freeman. *Dialectics and the macrostructure of arguments: A theory of argument structure*, volume 10. Walter de Gruyter, 1991.

[Freeman, 2011] James B Freeman. *Argument Structure: Representation and Theory*. Springer, 2011.

[Ghosh *et al.*, 2014] Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, page 3948. Association for Computational Linguistic, 2014.

[Ghosh *et al.*, 2016] Debanjan Ghosh, Aquila Khanam, and Smaranda Muresan. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of ACL 2016*, 2016.

[Groarke, 1996] L. Groarke. Informal logic. $http://plato.stanford.edu/entries/logic-informal/$, 1996.

[Habernal and Gurevych, 2016a] Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, page (in press), 2016. Submission received: 2 April 2015; revised version received: 20 April 2016; accepted for publication: 14 June 2016. Pre-print available at http://arxiv.org/abs/1601.02403.

[Habernal and Gurevych, 2016b] Ivan Habernal and Iryna Gurevych. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page (to appear). Association for Computational Linguistics, November 2016.

[Habernal and Gurevych, 2016c] Ivan Habernal and Iryna Gurevych. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany, 2016. Association for Computational Linguistics.

[Habernal *et al.*, 2014] Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation mining on the web from information seeking perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014.*, 2014.

[Henkemans *et al.*, 2007] Francisca Snoeck Henkemans, Frans van Eemeren, and Peter Houtlosser. *Argumentative Indicators in Discourse. A Pragma-Dialectical Study*. Dordrecht: Springer, 2007.

[Hirst *et al.*, 2014] Graeme Hirst, Vanessa Wei Feng, Christopher Cochrane, and Nona Naderi. Argumentation, ideology, and issue framing in parliamentary discourse. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014.*, 2014.

[Janier and Reed, 2015] Mathilde Janier and Chris Reed. Towards a theory of close analysis for dispute mediation discourse. *Argumentation*, 0.1007/s10503-015-9386-y, 2015.

[Janier *et al.*, 2014] Mathilde Janier, John Lawrence, and Chris Reed. Ova+: An argument analysis interface. In *Computational Models of Argument (COMMA)*, 2014.

[Janier *et al.*, 2015] Mathilde Janier, Mark Aakhus, Katarzyna Budzynska, and Chris Reed. Modeling argumentative activity in mediation with Inference Anchoring Theory: The case of impasse. In *European Conference on Argumentation (ECA)*, 2015.

[Kirschner *et al.*, 2015] Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, Denver, CO, June 2015. Association for Computational Linguistics.

[Knott, 1996] Alister Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh, 1996.

[Konat *et al.*, 2016] Barbara Konat, John Lawrence, Joonsuk Park, Katarzyna Budzynska, and Chris Reed. A corpus of argument networks: Using graph properties to analyse

divisive issues. In *Proc. of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, 2016.

[Kouylekov and Negri, 2010] M. Kouylekov and M. Negri. An open-source package for recognizing textual entailment. In *Procs of ACL 2010 System Demonstrations*, pages 42–47, 2010.

[Landis and Koch, 1977] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159174, 1977.

[Lawrence and Reed, 2015] John Lawrence and Chris Reed. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO, June 2015. Association for Computational Linguistics.

[Lawrence and Reed, 2016] J. Lawrence and C.A. Reed. Argument mining using argumentation scheme structures. In P. Baroni, M. Stede, and T. Gordon, editors, *Proceedings of the Sixth International Conference on Computational Models of Argument (COMMA 2016)*, Berlin, 2016. IOS Press.

[Lawrence *et al.*, 2012] John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. AIFdb: Infrastructure for the argument web. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 515–516, 2012.

[Lawrence *et al.*, 2014] John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[Lawrence *et al.*, 2015] J. Lawrence, M. Janier, and C. Reed. Working with open argument corpora. In *Proceedings of the 1st European Conference on Argumentation (ECA 2015)*, Lisbon, 2015. College Publications.

[Lawrence *et al.*, 2016] John Lawrence, Rory Duthie, Katarzyna Budzynska, and Chris Reed. Argument analytics. In *Computational Models of Argument (COMMA)*. IOS Press, 2016.

[Lawrence *et al.*, 2017] J. Lawrence, J. Park, B. Konat, K. Budzynska, C. Cardie, and C. Reed. Mining arguments for mapping controversies in online deliberative democracy and erulemaking. *ACM Transactions on Internet Technology*, 17(3), 2017.

[Levenshtein, 1966] V. I. Levenshtein. *Binary Codes Capable of Correcting Deletions, Insertions and Reversals*, volume 10:707. Soviet Physics Doklady, 1966.

[Levy *et al.*, 2014] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1489–1500, 2014.

[Levy *et al.*, 2015] Ran Levy, Liat Ein-Dor, Shay Hummel, Ruty Rinott, and Noam Slonim. TR9856: A multi-word term relatedness benchmark. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 419–424, 2015.

[Lippi and Torroni, 2015a] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2), 2015.

[Lippi and Torroni, 2015b] Marco Lippi and Paolo Torroni. Context-independent claim detection for argument mining. In Qiang Yang and Michael Wooldridge, editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 185–191. AAAI Press, 2015.

[Marco *et al.*, 2006] Chrysanne Di Marco, Donald D. Cowan, Peter Bray, H. Dominic Covvey, Vic Di Ciccio, Eduard H. Hovy, Joan Lipa, and Douglas W. Mulholland. A physician's authoring tool for generation of personalized health education in reconstructive surgery. In *Argumentation for Consumers of Healthcare, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-01, Stanford, California, USA, March 27-29, 2006*, pages 39–46. AAAI, 2006.

[Moens *et al.*, 2007] Marie-Francine Moens, Erik Boiy, Raquel M. Palau, and Chris Reed. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. ACM, 2007.

[Moens, 2013] Marie-Francine Moens. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *FIRE '13 Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation*, 2013.

[Monz and de Rijke, 2001] C. Monz and M. de Rijke. Light-weight entailment checking for computational semantics. In *Procs of Inference in Computational Semantics (ICoS-3)*, pages 59–72, 2001.

[Naderi and Hirst, 2015] Nona Naderi and Graeme Hirst. Argumentation mining in parliamentary discourse. In *Proceedings of the 15th Workshop on Computational Models of Natural Arguments (CMNA 2015)*, 2015.

[Nguyen and Litman, 2015] Huy Nguyen and Diane Litman. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, CO, June 2015. Association for Computational Linguistics.

[Nguyen and Litman, 2016] Huy Nguyen and Diane J. Litman. Improving argument mining in student essays by learning and exploiting argument indicators versus essay topics. In Zdravko Markov and Ingrid Russell, editors, *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2016, Key Largo, Florida, May 16-18, 2016.*, pages 485–490. AAAI Press, 2016.

[Palau and Moens, 2009] Raquel M. Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM, 2009.

[Park and Cardie, 2014a] Joonsuk Park and Claire Cardie. Assess: A tool for assessing the support structures of arguments in user comments. In *Proc. of the Fifth International Conference on Computational Models of Argument*, pages 473–474, 2014.

[Park and Cardie, 2014b] Joonsuk Park and Claire Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[Peldszus and Stede, 2013] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: a survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31, 2013.

[Peldszus and Stede, 2015] Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In *First European Conference on Argumentation: Argumentation and Reasoned Action*, 2015.

[Perelman and Olbrechts-Tyteca, 1958] Ch. Perelman and L. Olbrechts-Tyteca. *Traite de l'argumentation: La nouvelle rhetorique*. Paris: Presses Universitaires de France, 1958.

[Prakken, 2010] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1:93–124, 2010.

[Rahwan *et al.*, 2007] I. Rahwan, F. Zablith, and C. Reed. Laying the foundations for a World Wide Argument Web. *Artificial Intelligence*, 171(10-15):897–921, 2007.

[Reed *et al.*, 2008] Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. Language resources for studying argument. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC-2008)*, pages 91–100, Marrakech, 2008.

[Reed, 2006] Chris Reed. Preliminary results from an argument corpus. In Elona Miyares Bermdez and Leonel Ruiz Miyares, editors, *Linguistics in the twenty-first century*, pages 185–196. Cambridge Scholars Press, 2006.

[Rinott *et al.*, 2015] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 440–450, 2015.

[Roitman *et al.*, 2016] Haggai Roitman, Shay Hummel, Ella Rabinovich, Benjamin Sznajder, Noam Slonim, and Ehud Aharoni. On the retrieval of wikipedia articles containing claims on controversial topics. In Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao, editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 991–996. ACM, 2016.

[Saint-Dizier, 2012] Patrick Saint-Dizier. Processing natural language arguments with the <textcoop> platform. *Journal of Argument and Computation*, 3(1):49–82, 2012.

[Saint-Dizier, 2014] Patrick Saint-Dizier. *Challenges of Discourse processing: the case of technical documents*. Cambridge Scholars Publishing, 2014.

[Searle and Vanderveken, 1985] J. Searle and D. Vanderveken. *Foundations of Illocutionary Logic*. Cambridge University Press, 1985.

[Searle, 1969] J. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, New York, 1969.

[Sonntag and Stede, 2014] Jonathan Sonntag and Manfred Stede. GraPAT: a tool for graph annotations. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[Sridhar *et al.*, 2015] Dhanya Sridhar, James R. Foulds, Bert Huang, Lise Getoor, and Marilyn A. Walker. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 116–125. The Association for Computer Linguistics, 2015.

[Stab and Gurevych, 2014] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1501–1510, 2014.

[Stede, 2004] Manfred Stede. The Potsdam Commentary Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102, 2004.

[Swanson *et al.*, 2015] Reid Swanson, Brian Ecker, and Marilyn A. Walker. Argument mining: Extracting arguments from online dialogue. In *Proc. of 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–225, 2015.

[Teufel *et al.*, 1999] Simone Teufel, Jean Carletta, and Marie-Francine Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 110–117. Association for Computational Linguistics, 1999.

[Teufel *et al.*, 2009] Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502. Association for Computational Linguistics, 2009.

[Toulmin, 1958] Stephen Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.

[van Eemeren *et al.*, 2014] Frans H. van Eemeren, Bart Garssen, Eric C.W. Krabbe, A.Francisca Snoeck Henkemans, Bart Verheij, and Jean H.M. Wagemans. *Handbook of Argumentation Theory*. Springer, 2014.

[Villalba and Saint-Dizier, 2012] Maria Paz G. Villalba and Patrick Saint-Dizier. Some facets of argument mining for opinion analysis. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 23–34, 2012.

[Wacholder *et al.*, 2014] Nina Wacholder, Smaranda Muresan, Debanjan Ghosh, and Mark Aakhus. Annotating multiparty discourse: Challenges for agreement metrics. In *Proceedings of the 8th Linguistic Annotation Workshop, COLING*, pages 120–128, 2014.

[Wachsmuth *et al.*, 2015] Henning Wachsmuth, Johannes Kiesel, and Benno Stein. Sentiment flow - A general model of web review argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 601–611, 2015.

[Walker *et al.*, 2012] Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *LREC*, pages 812–817, 2012.

[Wyner *et al.*, 2016] Adam Wyner, Tom van Engers, and Anthony Hunter. Working on the argument pipeline: Through flow issues between natural language argument, instantiated arguments, and argumentation frameworks. *Argument & Computation*, 7:6989, 2016.

Katarzyna Budzynska
Centre for Argument Technology
Polish Academy of Sciences, Poland
& University of Dundee, UK
Email: budzynska.argdiap@gmail.com

Serena Villata
Université Cte dAzur, CNRS, Inria, I3S
France
Email: villata@i3s.unice.fr