Argument Mining

Katarzyna Budzynska , Serena Villata

***Abstract*—Fast, automatic processing of texts posted on the Internet to find positive and negative attitudes towards products and companies gave *sentiment analysis*, an area of text mining, a significant application in predicting trends on stock markets. *Opinion mining* further extended the scope of the search to help companies, such as those specialising in media analysis, to automate extraction of people's beliefs about products, institutions, politicians, celebrities. Now, *argument mining* goes one more step ahead to provide us with instant information not only about what attitudes and opinions people hold, but also about arguments which people give in favour (pro) and against (con) these attitudes and opinions. When this rapidly developing technology will mature, it will allow us to automatically and empirically explore vast amount of social media data (rather than seeking advices and opinions of experts) to give us answers such as why people decided to vote for one presidential candidate rather than the other.**

***Index Terms*—Argumentation, debating, computational linguistics, text mining.**

I. Introduction

ARGUMENT mining (also referred to or associated with argumentation mining, computational argumentation or debating technologies) is a new and rapidly growing area of natural language processing, and more specifically – text mining, which both are disciplines belonging to computational linguistics and artificial intelligence (see e.g., [28], [31], [26], [5] for a more detailed overview). Its goal is to develop methods and techniques which allow for automatic identification and extraction of argument data from large resources of natural language texts.

The broad area of text mining aims to provide robust tools, methods and techniques which allow for speeding up processing, interpreting and making sense out of the large amount of datasets of texts in natural language. The growth of this area is driven by a problem of the explosion of data available on the Internet. While having vast amount of data is an unquestionable value, the resources become of limited usefulness if we can not process them efficiently in a relatively short time and with low cost. If a company, such as Amazon or eBay, receives a lot of feedback from customers, but it takes months to analyse reviews posted on the company's webpage during just one day, then such a feedback will have very limited use for the company to understand what people like or dislike about their products and service. An extreme of this problem is referred to as Big Data, i.e. a situation when data is produced faster than users of these data and standard computational methods can process them.

Argument mining is a natural continuation and evolution of sentiment analysis and opinion mining – two areas of text mining which became very successful and important both academically and commercially. In sentiment analysis, the work focuses on extracting people's attitudes (positive, neutral, negative) towards persons, events or products. One commercially successful application of this research area is stock market where it is possible to relatively quickly process vast amount of resources such as news and social media to extract information about trends and tendencies on the market and to predict changes in stock prices. In opinion mining, the work aims to mine people's opinions about persons, events or products, e.g. the opinion that UK economy will be stronger without contributing a vast amount of money to the EU budget or the opinion that the UK economy will be weakened without the access to the common EU market. Its main commercial application is media analysis which monitors media to identify people's reactions for new products, companies, presidential candidates and so on. Argument mining, on the other hand, allows for recognising not only *what* attitudes and opinions people hold, but also *why* they hold them.

The growth of the commercial interests in the area of argument mining is manifested through the involvement of companies in several academic projects as well as the development of techniques such as IBM's Watson Debater (see e.g., www.arg.tech/ibmdebater) which searches for arguments pro and con regarding a given topic in Wikipedia articles.

II. Pipeline of natural language processing techniques applied to argument mining

Argument mining pipeline comprises of linguistic and computational part (see Figure 1). 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 argument mining pipeline aims to develop grammars (structural approach) and classifiers (statistical approach) to automatically annotate 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) in order to provide as an output only these information which are valuable for us, i.e. structured argument data.

Katarzyna Budzynska is an associate professor (senior lecturer) in the Institute of Philosophy and Sociology of the Polish Academy of Sciences, Poland & a lecturer and Dundee fellow in the School of Computing at the University of Dundee, UK, e-mail: budzynska.argdiap@gmail.com (see www.argdiap.pl/budzynska/). Together with Professor Chris Reed, she runs the ***Center for Argument Technology*** (see www.arg.tech).

Serena Villata is a researcher (CR1) at the Centre National de la Recherche Scientifique (CNRS) in the I3S Laboratory, France, e-mail:villata@i3s.unice.fr (see www.i3s.unice.fr/˜villata/).

This report is a short version of our article "Processing Argumentation in Natural Language Texts" which will appear in *Handbook of Formal Argumentation* in 2017 [5].

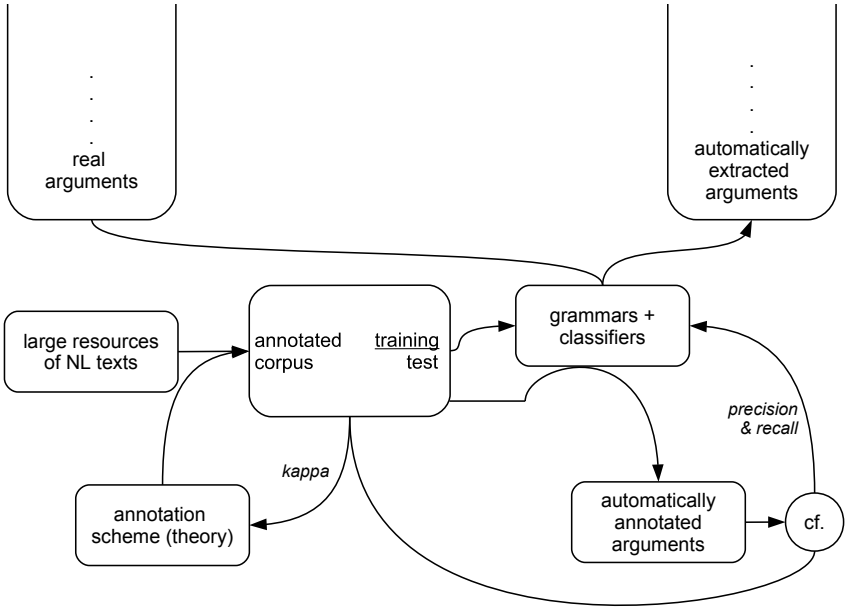

Fig. 1. A pipeline of natural language processing techniques applied to argument mining.

A. Databases of texts in natural language

The first step of the linguistic part of the pipeline starts with the task of *collecting large resources of natural language texts* (see "large resources of NL text" box in Figure 1) which then can be used for training and testing of argument mining algorithms. For example, Palau and Moens used dataset consisting of 92,190 words, 2,571 sentences divided in 47 documents from the European Court of Human Rights [27]; Habernal and Gurevych collected database comprising of 90,000 words in 340 documents of user-generated web discourse [15]; Garcia-Villalba and Saint-Dizier used 21,500 words in 50 texts as a test corpus [39].

Typically, the task of argument mining is narrowed down to the specific type of discourse (genre), since algorithms use the linguistic surface for argument recognition with none or little knowledge about the world, discourse context or deeper pragmatic level of a text. Genres studied up to date range from legal texts (e.g., [27], [2]); mediation (e.g., [19]); scientific papers (e.g., [38], [20]); online comments (e.g., [39], [30], [14], [40]); political debates (e.g., [18], [10]); technical texts (e.g., [33]); online debates (e.g., [41], [6], [35], [3], [16]); persuasive essays (e.g., [36], [13]); to Wikipedia articles (e.g., [1], [25]).

B. Theories & annotation schemes

The next step of argument mining pipeline consists of *choosing a model of argumentation which is then used to develop an annotation scheme* for analysing arguments in natural language texts. An annotation scheme for argumentative texts is a set of labels (tags) which defines arguments and their aspects for annotators (analysts) to use for structuring the dataset.

In the literature, there is a variety of different annotation schemes which aim to 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 given genre). In one of the first work in the argument mining [27], Palau and Moens choose a basic, intuitive conceptualisation of argument structure which consists of three labels: (a) premise: statements which provides a support; (b) conclusion: statements which are supported; (c) argument: a full structure comprising of premises and conclusion.

In her Argumentative Zoning work [38], Teufel uses more complex set of labels specifically tailored for mining argumentation in scientific texts: (a) background: general scientific background; (b) other: neutral descriptions of other people's work; (c) own: neutral descriptions of the own, new work; (d) aim: statements of the particular aim of the current paper; (e) textual: statements of textual organization of the current paper (e.g. "In chapter 1, we introduce..."); (f) contrast: contrastive or comparative statements about other work; explicit mention of weaknesses of other work; and (g) basis: statements that own work is based on other work.

Peldszus and Stede [31] introduce an annotation scheme drawing on different ideas from the literature and their practical experiences with analysing texts in the Potsdam Commentary Corpus [37]. The schema follows Freeman's idea of using the moves of proponent and challenger in a basic dialectical situation as a model of argument structure [12] with the representation of the rebutting/undercutting distinction and complex attack- and counter-attack constellations. Their scheme considers five kinds of supports among premises and the claim: (a) basic argument, (b) linked support, (c) multiple support, (d) serial support, and (e) example support; four kinds of challenger's attacks of the proponent's argument: (a) rebut a conclusion, (b) rebut a premise, (c) undercut an argument, (d) and support of a rebutter; and four proponent's counter-

attacks of the challenger's attack: (a) rebut a rebutter, (b) rebut an undercutter, (c) undercut a rebutter, and (d) undercut an undercutter.

An annotation scheme which considers the broad dialogical context of argumentation was proposed in [4]. Building upon Inference Anchoring Theory, Budzynska *et al.* extend the set of tags for arguments pro and con with dialogue structures and illocutionary structures [34] with two groups of tags. For the MM2012 corpus (www.corpora.aifdb.org/mm2012), the annotators could use the following tags associated with individual moves of a speaker in the dialogue: (a) asserting, (b) questioning (pure, assertive, and rhetorical), (c) challenging (pure, assertive, and rhetorical), and (d) popular conceding (statement that is assumed to belong to general knowledge); and for tags associated with the interactions between speaker(s)' moves in the dialogue, the annotators could choose between: (a) agreeing, (b) disagreeing, and (c) arguing.

C. Manual annotation & corpora

The *process of annotation* starts with segmenting (splitting) the text into elementary discourse units (EDUs) or in fact into argumentative discourse units (ADUs). Annotators use software tools such as the `arggraph` DTD[1], the RSTTool[2], the Glozz annotation tool[3] and OVA+[4], which help them to assign labels from the annotation schemeset to ADUs directly in a code.

Next, the annotated data have to be stored as a corpus. For example, the IBM Debating Technologies corpus[5] contains three different datasets: the dataset for automatic detection of claims and evidence in the context of controversial topics (1,392 labeled claims for 33 different topics) [1], and its extended version (2,294 labeled claims and 4,690 labeled evidence for 58 different topics). Another 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 topic, stance, degree of agreement, sarcasm, and nastiness among others [41]. The UKPConvArg1[6] corpus is a recently released dataset composed of 16,000 pairs of arguments over 32 topics annotated with the relation "A is more convincing than B" [16].

As the manual annotation is a highly time-consuming task, sharing and reusing analysed data becomes a real value. This is an objective of the freely accessible database AIFdb (www.aifdb.org) [24] which hosts multiple corpora (www.corpora.aifdb.org, see Figure 2). The key advantage of AIFdb is that it uses a standard for argument representation – the Argument Interchange Format, AIF [7]. The corpora 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 by the group in Santa Cruz [41]. Currently this database has

[1] www.github.com/peldszus/arg-microtexts/blob/master/corpus/arggraph.dtd
[2] www.wagsoft.com/RSTTool/
[3] www.glozz.org
[4] www.ova.arg-tech.org/
[5] www.researchweb.watson.ibm.com/haifa/dept/vst/mlta_data.shtml
[6] www.github.com/UKPLab/acl2016-convincing-arguments

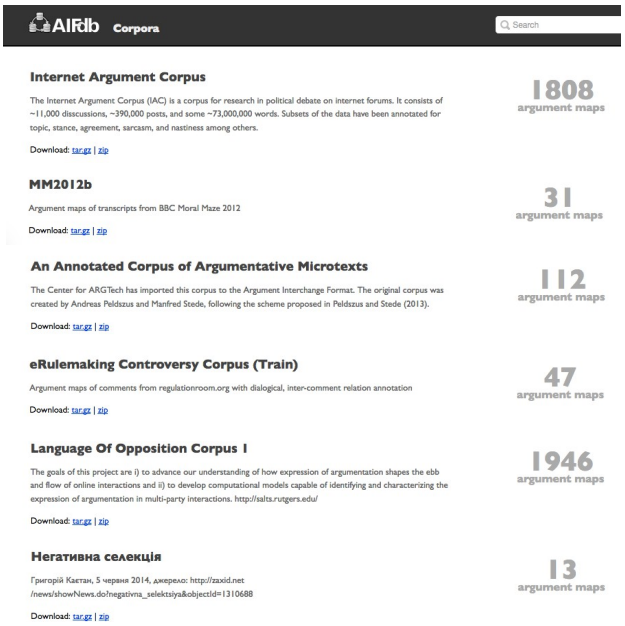

Fig. 2. Freely available AIFdb corpora.

300-500 unique users per month; stores 1,600,000 words and almost 57,000 annotated arguments in 15 languages (statistics obtained in November 2016).

D. Evaluation of manual step of annotation

The last step of the linguistic part of the argument mining pipeline is the *evaluation of the quality of a manual annotation* for which two comparison measures are the most typically used: (a) simple agreement which calculates a proportion (percentage) of matches between the analyses delivered by two annotators; or (b) several different kappa κ measures. The first one does not take into account the possibility of random matches, as if the annotators were tossing a coin and then assig labels according to the result. Thus, κ measures was introduced, amongst which the most popular one – Cohen's kappa [8] – shows the agreement between two annotators who each classify N items (e.g., ADUs) into C mutually exclusive categories (tags):

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

where $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the hypothetical probability of chance agreement.

The following scale [22] aims to interpret the level of agreement: 0.41-0.6 means moderate agreement, 0.6-0.8 is treated as substantial agreement, and 0.81-1 is assumed to be almost perfect agreement.

Recently, Duthie *et al.* proposed a new CASS metric, Combined Argument Similarity Score [9], which helps to avoid double penalising if the analysis involves different levels such as both segmentation and identification of argument structure. Arguments do not always span full sentences and automatic solutions may miss some tokens, this can then have a knock

on effect on the argumentative or dialogical structure with text spans being either larger or smaller and the κ penalising for this twice.

As an example, in eRulemaking corpus [29], the inter-annotator agreement was measured on 30% of the data resulting in Cohens κ of 0.73; in the MM2012 corpus [4], kappa for three types of illocutionary connections (arguing, agreeing and disagreeing) was $\kappa = 0.76$; in the persuasive essays corpus [36] inter-annotator agreement was measured on 90 persuasive essays for three annotators resulting in a Krippendorff's inter-rater agreement of α = 0.81[7]; and in the argumentative microtexts corpus [32] three annotators achieved an agreement of Fleiss κ = 0.83[8] for the full task.

E. NLP techniques

The next step *moves us to the computational part of the argument mining pipeline* (see "grammars + classifiers" box in Figure 1). In principle, there are two basic styles of automation (in practice, they are often combined to form a hybrid approach): (a) the structural approach, i.e. grammars (hand coded set of rules); and (b) the statistical approach, i.e. machine learning (general learning algorithms).

In the structural approach, a linguist looks through a selected fragment of a corpus (training corpus which in this case is more often referred to as a development corpus) and aims to find patterns between different lexical cues in the text and categories in the annotation scheme. For instance, in a given corpus it might be observed that arguments are linguistically signalled by words such as "because", "since", "therefore". Then, the linguist formulates rules describing these patterns in a grammar. The statistical approach 'replaces' a linguist with an algorithm. In the same way as a human, a system will also look for patterns, however, this time statistically on a larger sample of the training corpus.

A lot of work in argument mining applies the typical, 'off the shelf' NLP methods and techniques which are then further enriched to adapt them to a specific domain or genre of argumentative texts. Apart from discourse indicators such as "because", "since", "therefore" (see e.g., [21], [17]), different projects employ various additional information to improve the searching process for arguments such as e.g., argumentation schemes [11], the dialogical context [4], and the semantic context [6], or combination of different cues and techniques.

An example of structural approach is the work by Garcia-Villalba and Saint-Dizier [39] who investigate how an automatic recognition of arguments can be implemented in the Dislog programming language on the <TextCoop> discourse processing platform, or more precisely – whether argument mining techniques allows for capturing consumers' motivations expressed in reviews why they like or dislike a product. For instance, 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, which 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 majority of the work in argument mining employs, however, the statistical approach. Among them, Lippi and Torroni [25] present a framework to detect claims in unstructured corpora without necessity of resorting to contextual information. Their methodology is driven by the observation that argumentative sentences are often characterized by common rhetorical structures. As the structure of a sentence could be highly informative for argument detection, and in particular for the identification of a claim, the authors choose constituency parse trees for representing such information. They therefore build a claim detection system based on a Support Vector Machine (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. Habernal and Gurevych [16] aim to assess qualitative properties of the arguments to explain why one argument is more convincing than the other one. Based on a corpus of 26,000 annotated explanations written in natural language, two tasks are proposed on this data set, i.e., the prediction of the full label distribution; and the classification of the types of flaws in less convincing arguments. Cabrio and Villata [6] propose a framework to predict the relations among arguments using textual entailment (TE), a generic framework for applied semantics, where linguistic objects are mapped by means of semantic inferences at a textual level. TE is then coupled together with an abstract bipolar argumentation system which allows to identify the arguments that are accepted in online debates. The accuracy of this approach in identifying the relations among the arguments in a debate is about 75%.

F. Automatically annotated data

A system developed in the NLP stage is then used to *process raw, unannotated text in order to automatically extract arguments*. These texts have to be the same as the set of texts which was manually annotated and stored as a test corpus (see Figure 1). This step can be treated as an automated equivalent for manual annotation and corpus development.

Figure 3 shows an example of the output of a software tool. The <TextCoop> platform produces automatic segmentation and annotation. The text is split into argumentative discourse units (ADUs) which contain a minimal meaningful building blocks of a discourse with argumentative function. These propositional contents are presented as text in purple. Then, the system assigns illocutionary, communicative intentions (text in green) to ADUs of a type of assertions, rhetorical questions (RQ), and so on; as well as polymorphic types to represent the ambiguity (or underspecification) such as RQ-AQ which means that an ADU can be interpreted as having rhetorical questioning or assertive questioning illocution.

G. Evaluation of automatic step of annotation

The last step in the argument mining pipeline is the *evaluation of the quality of the automatic annotation*. A simple

[7]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.

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

```
< 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>.
```

Fig. 3. Example of data automatically annotated using the <TextCoop> platform for discourse processing of dialogical arguments in the MM2012 corpus.

measure, which is often used for this task, is accuracy, i.e. a proportion (percentage) of matches between manual and machine assignments of labels. If we want, however, to capture further, more detailed information about how well the system performed in mining arguments, a group of metrics: recall, precision and F_1 score, can be used. Let true positives, tp, will be a count how many times a machine assigned a label to the same text span as human analyst did; true negatives, tn – how often the machine did not assign a label to an ADU and the human did not either; false positives, fp – how often the machine assigned the label to a given text span while human did not; and false negatives, fn – how often the machine did not assign the label to a segment to which human made the assignment. Then:

– recall measures how many times the system did not recognise ("missed out") arguments:

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

– precision shows how many times the program found arguments correctly:

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

– 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 turns out to be not satisfactory, then we need to repeat the computational part of the process of argument mining trying to improve NLP techniques and methods we are using.

In their work, e.g., Palau and Moens obtain the following F_1 scores: 0.68 for the classification of premises; 0.74 for the classification of conclusions; and 0.6 for the determination of argument structures [27]. In [23], Lawrence and Reed aim to use argumentation schemes and combine different techniques in order to improve the success of recognising argument structure. This allows them to obtain the following results: for the technique of Discourse Indicators the system delivers precision of 1, recall of 0.08, and F_1 score of 0.15; for the technique of Topic Similarity the system has precision of 0.7, recall of 0.54 and F_1 score of 0.61; for the technique of Schematic Structure the system delivers precision of 0.82, recall of 0.69, and F_1 score of 0.75; and finally for the combination of these techniques the system improves the performance and delivers precision of 0.91, recall of 0.77, and F_1 score of 0.83.

III. Conclusion

This paper outlined the raising trends of the very recent argument mining research field. First of all, it is important to distinguish between the well-known NLP research field of *opinion mining* (or *sentiment analysis*) and argument mining. Besides 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 certain topic [14]. Second, argument mining approaches can support formal argumentation approaches to define formal models closer to human reasoning, where the fuzziness and ambiguity of natural language plays an important role and where the intellectual process is not always completely rational and objective. Actually, argument mining can provide more insights to answer questions like "what are the best arguments to influence a real audience?" and "what is the role of emotions in the argumentation process?".

As discussed also in the surveys of argument mining [31], [26], argument mining approaches face two main issues nowadays: big data and deep learning. Concerning the former, a huge amount of data is now available on the Web, such as social network posts, forums, blogs, product reviews, user comments to newspapers articles, and needs to be automatically analysed as it goes far beyond human capabilities to parse and understand it without any automatic support tool. Argument mining can make the difference here, and can exploit the Web to perform crowd-sourcing assessments to annotate very large corpora despite the difficulty of the task. Concerning the latter, deep learning methods, i.e., fast and efficient machine learning algorithms such as word embeddings, can be exploited in the argument mining pipeline to deal with large corpora and unsupervised learning.

Acknowledgment

Some research reported in this report was supported in part by EPSRC in the UK under grant EP/M506497/1 and in part by the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts".

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Katarzyna Budzynska and Serena Villata. *Handbook of Formal Argumentation*, chapter Processing Argumentation in Natural Language Texts. 2017, to appear.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

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

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

[23] 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.

[24] 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.

[25] 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.

[26] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10, 2016.

[27] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Law (ICAIL 2009),*, pages 98–109. ACM, 2009.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

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

[35] 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.

[36] 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.

[37] Manfred Stede. The potsdam commentary corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102, 2004.

[38] 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.

[39] 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.

[40] 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.

[41] 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.