Report from Dagstuhl Seminar 16161

Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments

Edited by

Elena Cabrio[1], Graeme Hirst[2], Serena Villata[3], and Adam Wyner[4]

1 **Laboratoire I3S – Sophia Antipolis, FR,** elena.cabrio@unice.fr
2 **University of Toronto, CA,** gh@cs.toronto.edu
3 **Laboratoire I3S – Sophia Antipolis, FR,** villata@i3s.unice.fr
4 **University of Aberdeen, UK,** azwyner@abdn.ac.uk

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 16161 "Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments", 17–22 April 2016. The seminar brought together leading researchers from computational linguistics, argumentation theory and cognitive psychology communities to discuss the obtained results and the future challenges of the recently born Argument Mining research area. 40 participants from 14 different countries took part in 7 sessions that included 30 talks, two tutorials, and a hands-on "unshared" task.

Seminar April 17–22, 2016 – http://www.dagstuhl.de/16161

1998 ACM Subject Classification I.2.4 Knowledge Representation Formalisms and Methods, I.2.7 Natural Language Processing

Keywords and phrases Argument Mining, Argumentation Theory, Cognitive Science, Computational Linguistics

Digital Object Identifier 10.4230/DagRep.6.4.80

Edited in cooperation with Alexis Palmer

1 Executive Summary

Graeme Hirst
Elena Cabrio
Serena Villata
Adam Wyner

Philosophers and, in more recent years, theorists working largely within Artificial Intelligence have developed formal computational models of argumentation, how it works, and what makes an argument valid or invalid. This work has made substantial progress in abstract, formal models to represent and reason over complex argumentation structures and inconsistent knowledge bases. Relatively little research, however, has applied these computational models to naturally occurring argumentation in text; nor have Computational Linguistics and Natural Language Processing substantially examined argumentation in text. Moreover, much of the work to date has studied only domain-specific texts and use-cases. Examples include finding the specific claims made in a scientific paper and distinguishing argumentation from narrative in legal texts.

Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments, *Dagstuhl Reports*, Vol. 6, Issue 4, pp. 80–109
Editors: Elena Cabrio, Graeme Hirst, Serena Villata, and Adam Wyner
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

But there are many uses and applications for automatic processing of the argumentative aspects of text, such as summarizing the argument of a complex court decision, helping a writer to structure an argument, and processing a large corpus of texts, such as blogs or consumer comments, to find arguments within it. To identify and integrate arguments across a corpus is a very significant problem. To address the issues, solve problems, and build applications, tools must be developed to analyze, aggregate, synthesize, structure, summarize, and reason about arguments in texts. Such tools would enable users to search for particular topics and their justifications, to trace through the argument, and to systematically and formally reason about the relations among arguments. However, to do so requires more linguistic sophistication and newer techniques than currently found in NLP. Moreover, NLP approaches must be connected to computational models of argument. The issues and problems have started to receive attention from both communities; for example, legal documents, on-line debates, product reviews, newspaper articles, court cases, scientific articles, and other kinds of text have all been the subject of recent NLP research on argumentation mining and have been tied to computational models.

Because argumentation is an inherently cross-disciplinary topic involving philosophy, psychology, communications studies, linguistics, and computer science, where different interpretations, analyses, and uses of arguments are proposed and applied, for progress in building NLP tools for argumentation there needs to be progress not only within each domain, but in bridging between these various disciplines, Natural Language Processing, and the computational models. This seminar aimed to help build this bridge by bringing together researchers from different disciplines, with the following goals:

- To understand better the specific kinds of tasks that NLP can carry out in argumentation.
- To establish a set of domain-specific and cross-domain use-cases that will guide the direction of research in the field.
- To understand better how computational argumentation tasks are tied – or not tied – to their specific domains, such as scientific papers, legal argumentation, and political discussions, looking for new cross-domain generalizations.
- To understand better the technical challenges to success in each of these tasks, and to discuss how the challenges can be addressed.
- To develop and explicate specific challenge problems for the integration of argumentation theory and NLP that are beyond the state of the art (but not too much so), and in which success would have the greatest effect on the field.
- To provide prototype solutions that address issues in the integration of NLP and argumentation theory, and to outline follow-on development.
- To propose or provide preliminary solutions to common open challenges in natural language argumentation (among others: argument retrieval in text, argument summarization, identification of semantic relations among arguments), profiting from the cross-fertilization between researchers coming from the research areas of NLP and formal argumentation.

The seminar was held on 17–22 April 2016, with 40 participants from 14 different countries. The event's seven sessions included 30 talks, two tutorials and a hands-on "unshared" task. The program included several plenary presentations and discussions in smaller working groups. The presentations addressed a variety of topics, as argument mining applied to legal argumentation and to writing support. Collective discussions were arranged for most of these topics, as well as plans for a future interdisciplinary research agenda involving experts from social sciences and psychology.

As a result of the seminar, a number of challenges and open issues have been highlighted:

- At this stage of maturity of the research area, it is difficult to choose good (possibly new) challenges and to define the task(s) to be addressed by automated systems

- Similarly, it is also challenging to precisely define and accomplish annotation task(s) to establish benchmarks and gold standards to test such automated systems
- It is essential to the fruitful development of the research area establish an Interdisciplinary outreach, involving social sciences, psychology, and economics.

Addressing these issues and other questions is now on the agenda of the Argument Mining research community.

2 Table of Contents

3 Overview of Talks

3.1 Putting Argument Mining to Work: an Experiment in Legal Argument Retrieval Using the LUIMA Type System and Pipeline

Kevin D. Ashley (University of Pittsburgh, US)

This paper highlights some results from an experiment demonstrating the feasibility of legal argument retrieval of case decisions in a particular legal domain. The LUIMA program annotates legal texts in terms of the roles some of the sentences play in a legal argument. It uses the annotations of argument-related information to re-rank cases returned by conventional legal information retrieval methods in order to improve the relevance of the top-ranked cases to users' queries. The experiment assessed the effectiveness of the re-ranking empirically and objectively, demonstrating how argument mining can be put to work.

3.2 Expert Stance Graphs for Computational Argumentation

Roy Bar-Haim, Noam Slonim, and Orith Tolego-Ronen

Main reference Will be presented in the 3rd Workshop on Argument Mining ACL 2016, Berlin.
URL http://argmining2016.arg.tech/

We describe the construction of an Expert Stance Graph, a novel, large-scale knowledge resource that encodes the stance of more than 100,000 experts towards a variety of controversial topics. We suggest that this graph may be valuable for various fundamental tasks in computational argumentation. Experts and topics in our graph are Wikipedia entries. Both automatic and semi-automatic methods for building the graph are explored, and manual assessment validates the high accuracy of the resulting graph.

3.3 Naïve arguments about argument mining

Pietro Baroni (University of Brescia, IT)

This discussion talk presents some questions about argument mining from an outsider perspective, with the goal of discussing some basic alternatives in approaching this important research problem, pointing out some aspects that may be worth clarifying to a newcomer to the field, and providing hints about what other research fields in the area of computational argumentation can offer. In particular the talk touches the following issues: holistic vs. restricted approaches, simple abstract vs. detailed argumentation models, domain dependence vs. context dependence, annotated corpora vs. paradigmatic examples, mining vs. generating arguments, and a priori vs. discovered ontologies.

3.4 Design Reasoning and Design Rationale – an experiment with implications for argument mining

Floris Bex (Utrecht University, NL), Rizkiyanto, and Jan Martijn van der Werf

In system and software design, we are faced with a combination of design reasoning, the process of analysing issues, options, pros and cons, and design rationale, the product of this process in the form of explicit issues, options, pros and cons. In our research we propose a basic card game aimed at improving the design reasoning process. This card game was used in an experiment, and the resulting design deliberations were transcribed and coded, where special care was taken to annotate the different design rationale elements posed by the participants. This, we can investigate the links between design reasoning phases and design rationale elements. In future research, the aim is to use process mining to analyse the design reasoning and, ultimately, try to (semi-)automatically mine design rationale from design reasoning discussions.

3.5 Interfaces to Formal Argumentation

Federico Cerutti (Cardiff University, GB)

Joint work of Federico Cerutti, Nava Tintarev, and Nir Oren
Main reference F. Cerutti, N. Tintarev, N. Oren, "Formal Arguments, Preferences, and Natural Language Interfaces to Humans: an Empirical Evaluation", in Proc. of the 21st European Conference on Artificial Intelligence (ECAI'14), Frontiers in Artificial Intelligence and Applications, Vol. 263, pp. 207–212, IOS Press, 2014.
URL http://dx.doi.org/10.3233/978-1-61499-419-0-207

Like other systems for automatic reasoning, argumentation approaches can suffer from "opacity." We explored one of the few mixed approaches explaining, in natural language, the structure of arguments to ensure an understanding of their acceptability status. In particular, we summarised the results described in [1], in which we assessed, by means of an experiment, the claim that computational models of argumentation provide support for complex decision making activities in part due to the close alignment between their semantics and human intuition. Results show a correspondence between the acceptability of arguments by human subjects and the justification status prescribed by the formal theory in the majority of the cases. However, post-hoc analyses show that there are some deviations. This seems to suggest that there is the need for some effort for making formal argumentation process more transparent.

References

1 Federico Cerutti, Nava Tintarev, and Nir Oren. *Formal Arguments, Preferences, and Natural Language Interfaces to Humans: an Empirical Evaluation*. In *21st European Conference on Artificial Intelligence*, pages 207–212, 2014.

3.6 Argument Extraction Challenges in a New Web Paradigm

Giorgos Flouris (FORTH – Heraklion, GR), Antonis Bikakis, Theodore Patkos, and Dimitris Plexousakis

The exchange of comments, opinions and arguments in blogs, social media, commercial websites or wikis is transforming the Web into a modern agora, a virtual place where all types of debates take place. This wealth of information remains unexploited: the purely textual form of online arguments leaves limited room for automated processing and the available methodologies of computational argumentation focus on logical arguments, failing to properly model online debates. We envision a formal, machine-interpretable representation of debates that would enable the discovery, tracking, retrieval, combination, interrelation, extraction and visualization of the vast variety of viewpoints that already exist on the Web, in a way that goes beyond simple keyword-based processing. This paper describes this vision and the related challenges, focusing on challenges related to argument extraction.

3.7 On Recognizing Argumentation Schemes in Formal Text Genres

Nancy L. Green (University of North Carolina – Greensboro, US)

Argumentation mining research should address the challenge of recognition of argumentation schemes in formal text genres such as scientific articles. This paper argues that identification of argumentation schemes differs from identification of other aspects of discourse such as argumentative zones and coherence relations. Argumentation schemes can be defined at a level of abstraction applicable across the natural sciences. There are useful applications of automatic argumentation scheme recognition. However, it is likely that inference-based techniques will be required.

3.8 Argumentative Writing Support: Structure Identification and Quality Assessment of Arguments

Iryna Gurevych (TU Darmstadt, DE) and Christian Stab

Main reference C. Stab, I. Gurevych, "Parsing Argumentation Structures in Persuasive Essays", arXiv:1604.07370v2 [cs.CL], 2016.
URL https://arxiv.org/abs/1604.07370v2

Argumentation is an omnipresent daily routine. We engage argumentation not only for making decisions or convincing an audience but also for drawing widely accepted conclusions and inferring novel knowledge. Good argumentation skills are crucial to learning itself. Consequently, argumentation constitutes an important part of education programs. With the emergence of the *Common Core Standard*, argumentative writing receives increasing attention. However, many students are still underprepared in writing well-reasoned arguments since teachers are not able to provide sufficient writing assignments in view of increasing class sizes and the load for reviewing argumentative texts. In addition, current *Intelligent Writing Systems* (IWS) are limited to feedback about spelling, grammar, mechanics, and discourse structures, and there is no system that provides feedback about written arguments. Novel *Natural Language Processing* (NLP) methods that analyze written arguments and pinpoint the weak points in argumentative discourse could bridge this gap.

In this talk, we highlighted the recent research projects on *Computational Argumentation* (CA) at the UKP-Lab. In particular, we provided an overview of the methods developed in the context of *Argumentative Writing Support* (AWS). We highlighted the results of an annotation study on argumentation structures, introduced an argumentation structure annotated corpus of persuasive essays, and presented a novel end-to-end *argumentation structure parser* for extracting micro-level argumentation structures. We introduced the architecture of the system and outlined the evaluation results. In addition, we introduced two novel tasks and our experimental results on *quality assessment* of natural language arguments. First, we presented our work on identifying myside bias in persuasive essays. Second, we presented an approach for identifying insufficiently supported arguments.

3.9 Crowdsourced and expert annotations for argument frame discovery

Graeme Hirst (University of Toronto, CA)

Joint work of Naderi, Nona; Hirst, Graeme

Introduction

Theoretical perspectives on framing and frames are diverse, but these theories converge in their conceptualization of framing as a communication process to present an object or an issue. According to Entman (1993):

> Framing involves selection and salience. To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described.

In parliamentary discourse, politicians use framing to highlight some aspect of an issue when they expound their beliefs and ideas through argumentation. For example, a politician speaking about immigration, regardless of which side of the issue they are on, might frame it as one of economics, or of multiculturalism, or of social justice.

The long-term goal of our work is to build computational models for the automatic analysis of issue-framing in argumentative political speech. In this paper, we address the short-term goal of annotating issue-framing in argumentative political speech.

Crowd-driven annotation?

We experimented with crowd-driven annotation of the Parliamentary debates to create training and test data. Such an annotation task requires the crowd members to have time to read through the speeches, to have knowledge of the issue at hand, and to have skill at identifying the correct frame. Our experiments used text from the debates of the Canadian

Parliament on two recent issues: The proposed federal gun registry (1995) and gay marriage (instituted in Canada in 2005 but subsequently re-debated in Parliament in 2005–6). We selected paragraphs from Parliamentary debates on these topics and manually (intuitively) derived a set of frames and topics for each side of issue. We then created a CrowdFlower task: Annotate each paragraph with its frames and topics. Multiple selections were permitted as was "none of the above". For the gun-registry topic, we had 1200 paragraphs and identified 22 frames. They included: gun ownership is a privilege, not a right; people need guns for protection; and a gun registry would reduce crime. For gay marriage, we had 659 paragraphs and 13 frames. They included: gay marriage is unnatural as it can't produce children; and marriage is a human right.

The results demonstrated that this task was too difficult and was not possible for crowd members. After 16 days, we had only 285 judgements on the gun-registry texts, of which just 17 were trusted, and similar results on the gay-marriage texts. We tried making the task easier by reducing the number of frames to nine, but this made little difference. Moreover, there was little agreement between the crowd members and some expert annotations that we made.

Experiments on frame identification

We therefore turned to a smaller, simpler dataset as training data: the ComArg corpus of online debates in English, annotated with a set of these frames (Boltužić & Šnajder 2014). Prior manual analysis of online debates yielded set of 'standard' arguments for specific topics – in effect, frames for the topics – one of which was gay marriage for which there were seven standard arguments (three pro and four con). We used these as training data for identifying frames in Canadian parliamentary debates on the same topic.

But we still needed annotated Parliamentary texts for evaluation, so instead of crowd-sourcing, we asked our expert colleagues to act as annotators for 400 selected paragraph-length texts (two annotators each) and 136 selected sentences (three annotators each). We used only texts for which at least two annotators were in agreement, leaving us with 366 paragraph-length texts and 121 sentences.

To classify each Parliamentary text by frame, we used its textual similarity to the Boltužić & Šnajder arguments. We tried three different vector-based representations:

- word2vec embeddings (300 dimensions) (Mikolov et al. 2013), summing the vectors for the words to get a vector for the complete sentence or paragraph.
- Syntax-based word embeddings (300 dimensions) (Wang et al. 2015), again summed.
- Skip-thought sentence vectors (4800 dimensions) (Kiros et al. 2015).

For measurements of textual similarity to the statement of the argument we tried both the cosine between the vectors and concatenation of absolute difference between the vectors with their component-wise product ("p&d"). We used these measurements as features in a multiclass support-vector machine. In addition, we also tried using known stance as a possible additional feature. Our baselines were the majority class and using a simple bag-of-words similarity with $tf \cdot idf$ weights.

We obtained better accuracy with the p&d similarity measure, but there was no consistent pattern in the results with respect to representation. For the paragraph-length texts, the majority-class baseline was an accuracy of 53.3%, the bag-of-words baseline was 71.0% accuracy, and the best accuracy achieved by a combination of our methods was 75.4%; most combinations gave results at or well below baseline accuracy. For the single sentences, the baselines were 33.0% and 52.8% respectively, and our best accuracy was 73.5%; most combinations gave results lower than that and only a few points above baseline. (Complete tables of results are given by Naderi and Hirst (2016).) Although, as noted, there was no consistent pattern in the results, we observed that the syntactically informed representation never gave good results, which was a surprise as one would have expected it to be superior at least to the purely lexical word2vec representation.

Conclusion

In this preliminary study, we examined annotation and automatic recognition of frames in political argumentative discourse. It is very preliminary work, using a noisy, simplistic representation of frames as text. We found that most of the information used by the classifier is lexical and that syntax is surprisingly unhelpful. We also found that annotation of the texts was too difficult for crowdsourcing, which severely limits the size of frame-annotated corpora that we can build. We will explore semi-supervised or unsupervised approaches as an alternative.

Acknowledgements. This work is supported by the Natural Sciences and Engineering Research Council of Canada and by the Social Sciences and Humanities Research Council. We thank Patricia Araujo Thaine, Krish Perumal, and Sara Scharf for their contributions to the annotation of parliamentary statements, and Tong Wang for sharing the syntactic embeddings. We also thank Tong Wang and Ryan Kiros for fruitful discussions.

References

1 Boltužić, F., Šnajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. *Proceedings of the First Workshop on Argumentation Mining*, pp. 49–58.
2 Entman, R.M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43(4), 51–58 (1993)
3 Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fidler, S. (2015). Skip-thought vectors. arXiv preprint arXiv:1506.06726.
4 Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
5 Naderi, Nona and Hirst, Graeme (2016). Argumentation mining in parliamentary discourse. To appear. (Available on request from the authors.)
6 Wang, Tong; Mohamed, Abdel-rahman; Hirst, Graeme (2015). Learning lexical embeddings with syntactic and lexicographic knowledge. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 458–463.

3.10 Aggregating evidence about the positive and negative effects of treatments

Anthony Hunter (University College London, GB) and Matthew Williams

Main reference A. Hunter, M. Williams, "Aggregating evidence about the positive and negative effects of treatments", Artificial Intelligence in Medicine, 56(3):173–190, 2012.
URL http://dx.doi.org/10.1016/j.artmed.2012.09.004

Objectives: Evidence-based decision making is becoming increasingly important in healthcare. Much valuable evidence is in the form of the results from clinical trials that compare

the relative merits of treatments. In this paper, we present a new framework for representing and synthesizing knowledge from clinical trials involving multiple outcome indicators.

Method: The framework generates and evaluates arguments for claiming that one treatment is superior, or equivalent, to another based on the available evidence. Evidence comes from randomized clinical trials, systematic reviews, meta-analyses, network analyses, etc. Preference criteria over argumentsare used that are based on the outcome indicators, and the magnitude of those outcome indicators, in the evidence. Meta-arguments attacks arguments that are based on weaker evidence.

Results: We evaluated the framework with respect to the aggregation of evidence undertaken in three published clinical guidelines that involve 56 items of evidence and 16 treatments. For each of the three guidelines, the treatment we identified as being superior using our method is a recommended treatment in the corresponding guideline.

Conclusions: The framework offers a formal approach to aggregating clinical evidence, taking into account subjective criteria such as preferences over outcome indicators. In the evaluation, the aggregations obtained showed a good correspondence with published clinical guidelines. Furthermore, preliminary computational studies indicate that the approach is viable for the size of evidence tables normally encountered in practice.

3.11 Introduction to Structured Argumentation

Anthony Hunter (University College London, GB)

Joint work of Besnard, Philippe; Hunter, Anthony
Main reference Ph. Besnard, A. Hunter, "Constructing Argument Graphs with Deductive Arguments: A Tutorial", Argument and Computation, 5(1):5–30, 2014.
URL http://dx.doi.org/10.1080/19462166.2013.869765

In this tutorial of structural argumentation, we focus on deductive argumentation. A deductive argument is a pair where the first item is a set of premises, the second item is a claim, and the premises entail the claim. This can be formalized by assuming a logical language for the premises and the claim, and logical entailment (or consequence relation) for showing that the claim follows from the premises. Examples of logics that can be used include classical logic, modal logic, description logic, temporal logic, and conditional logic. A counterargument for an argument A is an argument B where the claim of B contradicts the premises of A. Different choices of logic, and different choices for the precise definitions of argument and counterargument, give us a range of possibilities for formalizing deductive argumentation. Further options are available to us for choosing the arguments and counterarguments we put into an argument graph. If we are to construct an argument graph based on the arguments that can be constructed from a knowledgebase, then we can be exhaustive in including all arguments and counterarguments that can be constructed from the knowledgebase. But there are other options available to us. We consider some of the possibilities in this review [1].

References

1 Ph Besnard and A Hunter (2014) Constructing Argument Graphs with Deductive Arguments: A Tutorial, Argument and Computation, 5(1):5–30

3.12 Representing and Reasoning about Arguments Mined from Texts and Dialogues

Anthony Hunter (University College London, GB), Leila Amgoud, and Philippe Besnard

Main reference L. Amgoud, Ph. Besnard, A. Hunter, "Representing and Reasoning About Arguments Mined from Texts and Dialogues," in Proc. of the 13th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'15), LNCS, Vol. 9161, pp. 60–71, Springer, 2015.
URL http://dx.doi.org/10.1007/978-3-319-20807-7_6
Main reference L. Amgoud, Ph. Besnard, A. Hunter, "Logical Representation and Analysis for RC-Arguments", . in Proc. of the 27th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'15), pp. 104–110, IEEE CS, 2015.
URL http://dx.doi.org/10.1109/ICTAI.2015.28

This talk presents a target language for representing arguments mined from natural language. The key features are the connection between possible reasons and possible claims and recursive embedding of such connections. Given a base of these arguments and counterarguments mined from texts or dialogues, we want be able combine them, deconstruct them, and to analyse them (for instance to check whether the set is inconsistent). To address these needs, we propose a formal language for representing reasons and claims, and a framework for inferencing with the arguments and counterarguments in this formal language [1, 2].

References

1 Leila Amgoud, Philippe Besnard, Anthony Hunter: Representing and Reasoning About Arguments Mined from Texts and Dialogues. ECSQARU 2015: 60–71
2 Leila Amgoud, Philippe Besnard, Anthony Hunter: Logical Representation and Analysis for RC-Arguments. ICTAI 2015: 104–110

3.13 Working on the Argument Pipeline: Through Flow Issues between Natural Language Argument, Instantiated Arguments, and Argumentation Frameworks

Anthony Hunter (University College London, GB), Adam Wyner, and Tom van Engers

In many domains of public discourse such as arguments about public policy, there is an abundance of knowledge to store, query, and reason with. To use this knowledge, we must address two key general problems: first, the problem of the knowledge acquisition bottleneck between forms in which the knowledge is usually expressed, e.g. natural language, and forms which can be automatically processed; second, reasoning with the uncertainties and inconsistencies of the knowledge. Given such complexities, it is labour and knowledge intensive to conduct policy consultations, where participants contribute statements to the policy discourse. Yet, from such a consultation, we want to derive policy positions, where each position is a set of consistent statements, but where positions may be mutually inconsistent. To address these problems and support policy-making consultations, we consider recent automated techniques in natural language processing, instantiating arguments, and reasoning with the arguments in argumentation frameworks. We discuss application and "bridge" issues between these techniques, outlining a pipeline of technologies whereby: expressions in a controlled natural language are parsed and translated into a logic (a literals and rules knowledge base), from which we generate instantiated arguments and their relationships using a logic-based formalism (an argument knowledge base), which is then input to an implemented argumentation framework that calculates extensions of arguments (an argument extensions knowledge base), and finally, we extract consistent sets of expressions (policy positions). The paper reports progress towards reasoning with web-based, distributed, collaborative, incomplete, and inconsistent knowledge bases expressed in natural language.

References

1 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*, to appear, 2016.

3.14 Temporal Argument Mining for Writing Assistance

Diane J. Litman (University of Pittsburgh, US)

Joint work of Zhang, Fan; Litman, Diane K.

The written arguments of students are educational data that can be automatically mined for purposes of student instruction and assessment. While prior work has focused on argument mining within a single version of a text, our work focuses on temporal argument mining across a text and its revisions. This paper will illustrate some of the opportunities and challenges in temporal argument mining. I will briefly summarize how we are using natural processing to develop a temporal argument mining system, and how our system in turn is being embedded in an educational technology for providing writing assistance.

3.15 Locating and Extracting Key Components of Argumentation from Scholarly Scientific Writing

Robert Mercer (University of Western Ontario – London, CA)

Joint work of Graves, Heather; Graves, Roger; Ibn Faiz, Syeed; Houngbo, Hospice; Ansari, Shifta; Alliheedi, Mohammed; DiMarco, Chrysanne; Mercer, Robert

Mining the components of argumentation suggested by the Toulmin model from the text of scholarly scientific writings has been one research focus. Two such efforts are highlighted: examining titles as a source for the argumentative claim in experimental biomedical articles, and extracting higher order relations between two biomedical relations that correspond closely to warrants. The talk also presents an automated method to produce a sliver standard corpus for IMRaD sentence classification. Other sources of information pertaining to arguments are briefly introduced: (1) data in scientific writing is presented in tables and figures requiring non-linguistic methods to mine this information, (2) scientific arguments are not local to a single research article, so information from other articles could enhance the understanding of the argument in the article and place it in its broader scientific context, and (3) having a representation of how science is performed and then written about in the form of arguments could be beneficial.

References

1 Shifta Ansari, Robert E. Mercer, and Peter Rogan. 2013. Automated phenotype-genotype table understanding. In *Contemporary Challenges and Solutions in Applied Artificial Intelligence*, volume 489 of *Studies in Computational Intelligence*, pages 47–52. Springer.
2 Heather Graves, Roger Graves, Robert E. Mercer, and Mahzereen Akter. 2014. Titles that announce argumentative claims in biomedical research articles. In *Proceedings of the First Workshop on Argumentation Mining*, pages 98–99.
3 Hospice Houngbo and Robert E. Mercer. 2014. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 19–23.
4 Syeed Ibn Faiz and Robert E. Mercer. 2013. Identifying explicit discourse connectives in text. In *Proceedings of the 26th Canadian Conference on Artificial Intelligence (AI'2013)*, pages 64–76.
5 Syeed Ibn Faiz and Robert E. Mercer. 2014. Extracting higher order relations from biomedical text. In *Proceedings of the First Workshop on Argumentation Mining*, pages 100–101.
6 Mohammad Syeed Ibn Faiz. 2012. Discovering higher order relations from biomedical text. Master's thesis, University of Western Ontario, London, ON, Canada.
7 Barbara White. 2009. *Annotating a corpus of biomedical research texts: Two models of rhetorical analysis.* Ph.D. thesis, The University of Western Ontario, Canada.

3.16 Argumentation Mining in Online Interactions: Opportunities and Challenges

Smaranda Muresan (Columbia University – New York, US)

Joint work of Aakhus, Mark; Ghosh, Debanjan; Muresan, Smaranda; Wacholder, Nina
Main reference D. Ghosh, S. Muresan, N. Wacholder, M. Aakhus, M. Mitsui, "Analyzing Argumentative Discourse Units in Online Interactions", in Proc. of the First Workshop on Argumentation Mining at ACL, pp. 39–48, ACL, 2014.
URL http://acl2014.org/acl2014/W14-21/pdf/W14-2106.pdf

Argument mining of online interactions is in its infancy. One reason is the lack of annotated corpora in this genre. Another reason is that the coding of text as argument often misses how argument is an interactive, social process of reasoning. To make progress, we need to develop a principled and scalable way of determining which portions of texts are argumentative and what is the nature of argumentation. In this talk, I highlighted our approach to argumentation mining in online interactions that places a premium on identifying what is *targeted* and how it is *called out* (Ghosh et al., 2014; Wacholder et al., 2014; Aakhus, Muresan and Wacholder, 2013), and then I discussed some of the opportunities and challenges we face in this area.

Our approach defines an argumentative structure that contains the most basic components of interactive argumentation (i.e., CallOut, Target and Argumentative Relation) as well as finer-grained characteristics of CallOuts (e.g., Stance and Rationale components) and of Argumentative Relations (e.g., type Agree/Disagree/Other) (Ghosh et al., 2014; Wacholder et al., 2014). Our annotation study followed four desiderata: 1) Start with a coarse-grained argumentative structure (e.g., CallOut, Target, Argumentative Relations), and move to more fine-grained argumentative structures; 2) Code not just argument components (Target and CallOut) but also Argumentative Relations between these components; 3) Identify boundary points of argumentative discourse units (ADUs), which in principle can be of any length; and 4) Design an annotation granularity/scope that enables innovation in annotation by combining traditional and crowd-sourcing practices involving expert and novice annotators. We annotated a dataset of blog posts and their comments. Computationally, we tackled one specific problem: classifying the type of Argumentative Relation between a CallOut and a Target using local models (Ghosh et al., 2014).

The talk concluded with the discussion of several open issues:

- *Segmentation Step.* In our annotation study, the segmentation subtask (identifying argumentative discourse units), proved to be challenging since annotators were free to choose text spans of any length. First, our results show variation in the number of ADUs (e.g., CallOuts) identified by the expert annotators. The consistent variation among coders indicated that, beyond any potential training issues, annotators could be characterized as "lumpers", who treat a single long segment of text as one ADU and "splitters", who treat it as two (or more) shorter ADUs (Wacholder et al., 2014). Second, the segmentation variability meant that measurement of IAA had to account for fuzzy boundaries. Third, for developing computational models the segmentation step will be particularly challenging.
- *Computational Models.* Argument Relations are often long-distance and implicit. Most approaches to argumentation mining rely on local models. Recently, Peldszus and Stede (2015) proposed a global model that jointly predicts different aspects of the argument structure. A challenge will be to develop such global models based on discourse-parsing for online interactions, where we do not only have inter-sentence and intra-sentence relations, but also inter-turns relation.
- *Varied Datasets of Argumentative Texts.* As highlighted also by the Unshared Untask at the Dagstuhl seminar, there is a growing need for a common repository of different types of argumentative texts. For online interactions, our corpus consists of blog posts and their comments. Another source of online interactions is Reddit. A particularly interesting subreddit for argumentation mining is ChangeMyView, where posters are "people who have an opinion on something but accept that they may be wrong or want help changing their view." If a user is able to change someone else's view, they are awarded a delta point. Recently a large collection of ChangeMyView data has been released to the research community (Tan et al., 2016). An opportunity will be for working groups that focus on various aspects of argumentation to annotate the same datasets.

References

1 Ghosh, D., Muresan, S. , Wacholder, N. , Aakhus, M. and Mitsui, M. (2014). Analyzing argumentative discourse units in online interactions. *Proceedings of the First Workshop on Argumentation Mining*, 2014.

2 Aakhus, M., Muresan, S. and Wacholder, N (2013). Integrating natural language processing and pragmatic argumentation theories for argumentation support. *Proceedings of the 10th International Conference of the Ontario Society for the Study of Argumentation (OSSA 10): Virtues of Argumentation*, D. Mohammed and M. Lewiski (Eds.), 2013.

3 Wacholder, N., Muresan, S., Ghosh, D. and Aakhus, M. (2014). Annotating Multiparty Discourse: Challenges for Agreement Metrics. *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop*, 2014.

4 Peldszus, Andreas and Stede, Manfred (2015). Joint prediction in MST-style discourse parsing for argumentation mining. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

5 Tan, C., Niculae, V, Danescu-Niculescu-Mizil,C., and Lee, L. (2016). Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. *In Proceedings of the 25th International Conference on World Wide Web (WWW '16).*, 2016.

3.17 Argument Strength Scoring

Vincent Ng (University of Texas at Dallas, US)

Joint work of Persing, Isaac; Ng, Vincent
Main reference I. Persing, V. Ng, "Modeling Argument Strength in Student Essays," in Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing (Volume 1: Long Papers), pp. 543–552, ACL, 2015.
URL http://aclweb.org/anthology/P/P15/P15-1053.pdf

While recent years have seen a surge of interest in automated essay grading, including work on grading essays with respect to particular dimensions such as prompt adherence, coherence, and technical quality, there has been relatively little work on grading the essay dimension of argument strength. Argument strength, which refers to the persuasiveness of the argument an essay makes for its thesis, is arguably the most important aspect of argumentative essays. In this talk, I will introduce a corpus of argumentative student essays annotated with argument strength scores and propose a supervised, feature-rich approach to automatically scoring the essays along this dimension. I will conclude this talk with a discussion of the major challenges associated with argument strength scoring and argumentation mining in student essays.

3.18 Clause types in argumentative texts

Alexis M. Palmer (Universität Heidelberg, DE) and Maria Becker

Main reference M. Becker, A. Palmer,. A. Frank, "Argumentative texts and clause types", in Proc. of the 3rd Workshop on Argument Mining, pp. 21–30, ACL 2016.
URL http://aclweb.org/anthology/W/W16/W16-2803.pdf

This work is built on the theoretical framework of discourse mode theory, in which types of text passages are linked to linguistic characteristics of the clauses which compose the passages. [3] identifies five discourse modes: Narrative, Description, Report, Information, and Argument/Commentary. One way in which modes differ from one another is in their characteristic distributions of clause types. Taking the argumentative microtext corpus [2] as a set of prototypical argumentative text passages, we apply Smith's typology of clause types – know as situation entity (SE) types. The aim is to better understand what types of situations (states, events, generics, etc.) are most prevalent in argumentative texts and, further, to link this level of analysis to the argumentation graphs provided with the microtext corpus. The annotation project is ongoing, but preliminary analysis confirms that argumentative texts do in fact look different from non-argumentative texts with respect to SE types. This result suggests the potential for using systems for automatic prediction of SE types [1] to support computational mining of arguments from texts.

References

1 Annemarie Friedrich and Alexis Palmer and Manfred Pinkal. *Situation entity types: automatic classification of clause-level aspect*. In Proc. of ACL 2016. Berlin, Germany, 2016
2 Andreas Peldszus and Manfred Stede. *An annotated corpus of argumentative microtexts*. In Proceedings of the First European Conference on Argumentation: Argumentation and Reasoned Action. Lisbon, Portugal, 2015
3 Carlota S. Smith. *Modes of discourse: The local structure of texts*. Cambridge University Press, 2003

3.19 Joint prediction in MST-style discourse paring for argumentation mining

Andreas Peldszus (Universität Potsdam, DE) and Manfred Stede (Universität Potsdam, DE)

We introduce two datasets for argumentation mining: a selection of pro & contra newspaper commentaries taken from the German daily "Tagesspiegel", and a set of user-generated "microtexts". These have been produced in response to a trigger question such as "BLA", and consist of about 5 sentences, all of which are relevant for the argumentation. The microtext corpus consists of 112 texts, which have been translated from German to English. [5]

Our annotation scheme [3] used for both corpora builds on the proposals of [1] and constructs a full argument structure for a text: A central claim is (possibly recursively) backed by supporting statements. Following the inherently dialogical nature of argumentation, there may also be potential objections (by an "opponent") and accompanying counter-objections (by the "proponent"). Our experiments show that these can be reliably annotated by experts (kappa=0.83). In addition, we conducted experiments with lightly-trained students and employed clustering methods to identify reliable annotators, who also reached good agreement [2].

For automatic analysis of microtext argumentation, we developed an approach that divides structure prediction into four classification subtasks: central claim; statement's perspective as proponent or opponent; function as support or attack, and attachment between statements. We train individual local models for these, and then combine their predictions in a data structure we call "evidence graph", as it combines the individual contributions and allows for computing the globally preferred structure that respects certain well-formedness constraints. This process is implemented as minimal-spanning tree computation. It is guaranteed to yield complete and sound structures, and for the four subtasks, it gives significant improvements over the results of the local models [4].

Finally, we present an extension of the microtext corpus with two additional layers of discourse structure annotation: Rhetorical Structure Theory and Segmented Discourse Representation Theory. To achieve comparability, we first harmonized the segmentation decisions of the discourse layers and the argumentation layer. This allowed us to map all layers into a common dependency format based on identical discourse segments. To illustrate the potential of correlating the layers, we show how relations from RST correspond to those in the argumentation structure. In future work, on the one hand it is possible to systematically study the relationship between RST and SDRT. On the other hand, it will be fruitful to explore how argumentation mining can benefit from the presence of discourse structure information. [6]

References

1 Freeman, J. B.: Argument Structure: Representation and Theory. Argumentation Library (18). Springer, 2011.
2 Andreas Peldszus: Towards segment-based recognition of argumentation structure in short texts. Proceedings of the First Workshop on Argumentation Mining. ACL 2014. Baltimore, Maryland. pp. 88–97
3 Andreas Peldszus and Manfred Stede: From Argument Diagrams to Argumentation Mining in Texts: A Survey. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI) Vol. 7, No. 1, pp. 1–31. 2013

4 Andreas Peldszus and Manfred Stede: Joint prediction in MST-style discourse paring for argumentation mining. Proc. of EMNLP 2015, Lisbon.
5 Andreas Peldszus, Manfred Stede. An annotated corpus of argumentative microtexts. To appear in: First European Conference on Argumentation: Argumentation and Reasoned Action, Portugal, Lisbon, June 2015.
6 Manfred Stede, Stergos Afantenos, Andreas Peldszus, Nicholas Asher and Jérémy Perret: Parallel Discourse Annotations on a Corpus of Short Texts. Proc. of LREC 2016. Portoroz, Slovenia

3.20 Strategical Argumentative Agent for Human Persuasion

Ariel Rosenfeld (Bar-Ilan University – Ramat Gan, IL)

Main reference Strategical Argumentative Agent for Human Persuasion, 2016, Submitted.

Automated agents should be able to persuade people in the same way people persuade each other – via dialogs. Today, automated persuasion modeling and research use unnatural assumptions regarding the persuasive interaction which creates doubt regarding their applicability for real-world deployment with people. In this work we present a novel methodology for persuading people through argumentative dialogs. Our methodology combines theoretical argumentation modeling, machine learning and Markovian optimization techniques that together result in an innovative agent named SPA. Two extensive field experiments, with more than 100 human subjects, show that SPA is able to persuade people significantly more often than a baseline agent and no worse than people are able to persuade each other. This study is part of our ongoing effort to investigate the connections and challenges between Argumentation Theory and people [2, 1]. We hope that the encouraging results shown in this work (and in previous ones) will inspire other researchers in the field to investigate other argumentation-based methods in human experiments. We believe that bridging the gap between formal argumentation and human argumentation is essential for making argumentation practical for a wider range of applications.

References

1 A. Rosenfeld and S. Kraus. Providing arguments in discussions based on the prediction of human argumentative behavior. In *AAAI*, pages 1320–1327, 2015.
2 Ariel Rosenfeld and Sarit Kraus. 2014. Argumentation theory in the field: An empirical study of fundamental notions. In *Frontiers and Connections between Argumentation Theory and Natural Language Processing.*

3.21 Towards Knowledge-Driven Argument Mining

Patrick Saint-Dizier (CNRS-Paul Sabatier University – Toulouse, FR)

Main reference J., Pustejovsky, "The Generative Lexicon", MIT Press, 1995.
Main reference P. Saint-Dizier, "Processing natural language arguments with the TextCoop platform", Journal of Argumentation and Computation, 3(1), 2012.

Given a controversial issue, argument mining from texts in natural language is extremely challenging: besides linguistic aspects, domain knowledge is often required together with appropriate forms of inferences to identify arguments. This contribution explores the types of knowledge and reasoning schemes that are required and how they can be paired with language resources to accurately mine arguments. We show, via corpus analysis, that the Generative Lexicon (GL) structure, enhanced in different manners and associated with inferences and language patterns, is a relevant approach to capture the typical concepts found in arguments.

3.22 Medication safety as a use case for argumentation mining

Jodi Schneider (University of Pittsburgh, US) and Richard D. Boyce

URL http://www.slideshare.net/jodischneider/medication-safety-as-a-use-case-for-argumentation-mining-dagstuhl-seminar-16161-2016-0419

We present a use case for argumentation mining, from biomedical informatics, specifically from medication safety. Tens of thousands of preventable medical errors occur in the U.S. each year, due to limitations in the information available to clinicians. Current knowledge sources about potential drug-drug interactions (PDDIs) often fail to provide essential management recommendations and differ significantly in their coverage, accuracy, and agreement. The Drug Interaction Knowledge Base Project (Boyce, 2006-present; dikb.org) is addressing this problem.

Our current work is using knowledge representations and human annotation in order to represent clinically-relevant claims and evidence. Our data model incorporates an existing argumentation-focused ontology, the Micropublications Ontology. Further, to describe more specific information, such as the types of studies that allow inference of a particular type of claim, we are developing an evidence-focused ontology called DIDEO–Drug-drug Interaction and Drug-drug Interaction Evidence Ontology. On the curation side, we will describe how our research team is hand-extracting knowledge claims and evidence from the primary research literature, case reports, and FDA-approved drug labels for 65 drugs.

We think that medication safety could be an important domain for applying automatic argumentation mining in the future. In discussions at Dagstuhl, we would like to investigate how current argumentation mining techniques might be used to scale up this work. We can also discuss possible implications for representing evidence from other biomedical domains.

3.23 Social Media Argumentation Mining: The Quest for Deliberateness in Raucousness

Jan Šnajder (University of Zagreb, HR)

Joint work of Šnajder, Jan; Boltužić, Filip

Argumentation mining from social media content has attracted increasing attention. The task is both challenging and rewarding. The informal nature of user-generated content makes the task dauntingly difficult. On the other hand, the insights that could be gained by a large-scale analysis of social media argumentation make it a very worthwhile task. In this position paper I discuss the motivation for social media argumentation mining, as well as the tasks and challenges involved.

3.24 Assessing Argument Relevance at Web Scale

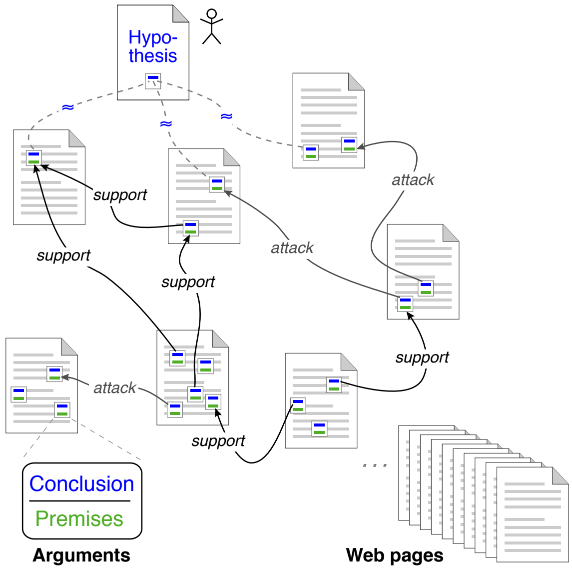
Benno Stein (Bauhaus-Universität Weimar, DE) and Henning Wachsmuth

Joint work of Stein, Benno; Wachsmuth, Henning

The information needs of users will focus more and more on arguments that can be found pro and con a queried hypothesis [3]. As a consequence, future information systems, above all web search engines, are expected to provide justifications for the results they return in response to user queries [7].

Accordingly, argument mining has become an emerging research topic, also being studied for the web [1]. Argument mining identifies the units of arguments (i.e., premises and conclusions) in natural language texts and it classifies their relations, but it does not clarify which arguments are *relevant* for a given hypothesis. First approaches to assess argument strength or similar exist [6, 2]. However, they hardly account for the fundamental problem that argument relevance is essentially subjective.

In our work, we envision the structural and hence objective assessment of argument relevance at web scale. To draw a clear line between existing work and the missing building blocks, we presume that technologies are available which can (1) robustly mine argument units from web pages and (2) decide if two units mean the same—or maybe the opposite. Based hereupon, we model all argument units and relations found on the web in an argument graph. We devise an adaptation of the famous PageRank algorithm [5], which recursively processes the argument graph to compute a score for each argument unit. From these scores, the relevance of arguments can be derived. The depicted figure below sketches a small argument graph. If we interpret the hypothesis of a user as a conclusion—as shown—all arguments with that conclusion may be relevant for the user's information need.



Originally, PageRank aims to measure the objective relevance of a web page based on all other pages that link to that page. Similarly, for an argument, we measure its relevance based on all other arguments that make use of its conclusion. Thereby, we separate conclusion relevance from the soundness of the inference an argument makes to arrive at its conclusion. In analogy to the supportive nature of web links, we restrict our "PageRank for argument relevance" to supporting argument units here. However, a variant based on attack relations is conceivable as well, covering ideas from [4] for the web then.

Practically, the construction of a reliable argument graph raises complex challenges of processing natural language texts. Also, the adapted PageRank brings up new questions, e.g., whether and how to balance support and attack. Following our approach, however, these questions can be addressed stepwise to bring argument relevance to search engines, starting from the technologies of today.

References

1 Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. *Cross-Domain Mining of Argumentative Text through Distant Supervision*. In Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics, pages 1395–1404, 2016.

2 Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. *Supporting Human Answers for Advice-Seeking Questions in CQA Sites*. In Proceedings of the 38th European Conference on IR Research, pages 129–141, 2016.

3 Elena Cabrio and Serena Villata. *Natural Language Arguments: A Combined Approach*. In Proceedings of the 20th European Conference on Artificial Intelligence, pages 205–210, 2012.

4 Phan Minh Dung. *On the Acceptability of Arguments and its Fundamental Role in Non-monotonic Reasoning, Logic Programming and n-Person Games*. Artificial Intelligence, 77(2):321–357, 1995.

5 Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66, Stanford InfoLab, 1999.

6 Isaac Persing and Vincent Ng. *Modeling Argument Strength in Student Essays*. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and

the 7th International Joint Conference on Natural Language Processing: Volume 1: Long Papers, pages 543–552, 2015.
7 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, pages 440–450, 2015.

3.25 Towards Relation-based Argumentation Mining?

Francesca Toni (Imperial College London, GB)

Joint work of Lucas Carstens
Main reference L. Carstens, F. Toni, "Towards relation based argumentation mining", in Proc. of the 2nd Workshop on Argumentation Mining, affiliated with NAACL 2015, pp. 29–34, ACL, 2015.
URL http://aclweb.org/anthology/W/W15/W15-0504.pdf

In this talk I overviewed foundations, tools and applications of Structured argumentation (ABA) and Abstract Argumentation (AA) for rule-based arguments as well as Bipolar argumentation and Quantitative Argumentation Debates (QuADs), see [1] for an overview. These frameworks can be supported by and support the mining of attack/support/neither relations amongst arguments (e.g. as in [2]). Moreover, I have disucssed the following questions: is the use of quantitative measures of strength of arguments, as proposed e.g. in QuADs, a good way to assess the dialectical strength of mined arguments or the doodness of argument mining? Can argumentation help argument mining?

References

1 Lucas Carstens, Xiuyi Fan, Yang Gao, and Francesca Toni. An overview of argumentation frameworks for decision support. In Madalina Croitoru, Pierre Marquis, Sebastian Rudolph, and Gem Stapleton, editors, Graph Structures for Knowledge Representation and Reasoning – 4th International Workshop, GKR 2015, Buenos Aires, Argentina, July 25, 2015, Revised Selected Papers, volume 9501 of Lecture Notes in Computer Science, pages 32–49. Springer, 2015.
2 Lucas Carstens and Francesca Toni. Towards relation based argumentation mining. In The 2nd Workshop on Argumentation Mining, affiliated with NAACL 2015, 2015.

3.26 The Need for Annotated Corpora from Legal Documents, and for (Human) Protocols for Creating Them: The Attribution Problem

Vern R. Walker (Hofstra University – Hempstead, US)

This presentation argues that in order to make progress today in automating argumentation mining from legal documents, we have a critical need for two things. First, we need a sufficient supply of manually annotated corpora, as well as theoretical and experimental evidence that those annotated data are accurate. Second, we need protocols for effectively training people to perform the tasks and sub-tasks required to create those annotations. Such protocols are necessary not only for a team approach to annotation and for quality assurance of the finished annotations, but also for developing and testing software to assist humans in the process of annotation. Drawing upon the latest work at Hofstra University's Law, Logic and Technology Research Laboratory in New York, the paper offers an extended example from the problem of annotating attribution relations, as an illustration of why obtaining consistent and accurate annotations in law is extremely difficult, and of why protocols are necessary. Attribution is the problem of determining which actor believes, asserts, or relies upon the truth of a proposition as a premise or a conclusion of an argument. The paper illustrates that in applying argumentation mining to legal documents, annotating attribution relations correctly is a critical task.

4 Working groups

4.1 A Pilot Study in Mining Argumentation Frameworks from Online Debates

Federico Cerutti (Cardiff University, GB), Alexis M. Palmer (Universität Heidelberg, DE), Ariel Rosenfeld (Bar-Ilan University – Ramat Gan, IL), Jan Šnajder (University of Zagreb, HR), and Francesca Toni (Imperial College London, GB)

We describe a pilot study mapping an online debate onto several types of Argumentation Frameworks, as understood in the AI field of Computational Argumentation. The pilot study aims to explore the richness of online debates *and* of Computational Argumentation methods and techniques. Additionally we consider the potential benefits of connecting the output of Argument Mining and the tools offered by Computational Argumentation, in particular algorithms in existing Argumentation Frameworks for determining the dialectical acceptability or strength of arguments. The mapping makes use of an intermediate graphical representation of the debate, manually generated by human annotators.

4.2 An Unshared Untask for Argumentation Mining

Ivan Habernal (TU Darmstadt, DE) and Adam Wyner

Introduction

In this extended abstract, we outline the *Unshared Untask* at the Argument Mining seminar at Dagstuhl. Argument mining (also known as "Argumentation mining") is a recent challenge in corpus-based discourse processing that applies a certain argumentation theory to model and automatically analyze the data at hand. Interest in Argument Mining has rapidly increased over the past few years, as demonstrated by a number of events (The BiCi seminar Frontiers and Connections between Argumentation Theory and Natural Language Processing in 2014; Workshops on Argumentation Mining at ACL 2014, NAACL 2015, and ACL 2016; the Dagstuhl Seminar on Debating Technologies in 2015; and the Dagstuhl seminar reported here on Natural Language Argumentation, to name a few). Given the wide the range of different perspectives and approaches within the community, it is now very important to consolidate the view and shape the further development of research on Argument Mining.

Shared tasks have become a major driver in boosting research in many NLP fields.[1] The availability of shared annotated data, clear evaluation criteria, and visible competing systems allow for fair, exact comparison between systems and overall progress tracking, which in turn fosters future research. However, argument mining, as an evolving research field, suffers not only from the lack of large data sets but also from the absence of a unified perspective on what tasks the systems should fulfill and how the systems should be evaluated. The tasks and evaluation measures require agreement, which is particularly problematic given the variety of argumentation models, argumentative genres and registers, granularity (e.g., micro-argumentation and macro-argumentation), dimensions of argument (logos, pathos, ethos), and the overall social context of persuasion.

The concept of a so-called "unshared untask" is an alternative to shared tasks. In an unshared untask, neither a clearly defined problem to be solved nor quantitative performance measures are given. Instead, participants are given only a variety of raw unannotated data and an open-ended prompt. The goals are to explore possible tasks, try to provide a rigorous definition, propose an annotation methodology, and define evaluation metrics, among others. This type of activity has been successful in several areas, such as PoliInformatics (a broad group of computer scientists, political scientists, economists, and communications scholars) [2] and clinical psychology and NLP [1]. Some venues also run shared and unshared tasks in parallel [3]. The nature of an unshared untask is mainly exploratory and can eventually lead to a deeper understanding of the matter and to laying down foundations for a future standard shared task.

The remainder of this extended abstract outlines the process, data, and observations about this activity.

Process

For this Dagstuhl seminar, we collected several diverse, relatively small samples as the basis for individual analysis and group discussion. We did not set any agenda or framework around analysis or outcome; rather, the individuals and groups were free to contribute what and as they thought appropriate. The objective was to generate a broad spectrum of ideas about the issues and challenges as well as the approaches and techniques.

In order to keep the 'idea pots stirring' in some organized fashion, we had created two 'rings' of three groups, and each ring worked with different datasets. Each group in a ring processed all the datasets in its ring over the course of three meetings.

For the closing panel discussion, we prepared and elicited critical questions to assess multiple aspects of the proposed tasks with respect to the corresponding data. This session summarized main observations and revealed possible future directions for the community, as will be presented later in this paper.

Data

We prepared the following data samples in a 'raw' form; the source texts were transformed into a plain-text format with only minimal structure kept (such as IDs of comments in Web

[1] See the various tracks in TREC http://trec.nist.gov/

discussions to properly identify which post a user is responding to). All visual cues (images, different font sizes) or non-related texts (HTML boiler-plate) were removed.

Scientific Paper Reviews contains three reviews and author responses from the NIPS 2013-2014 conferences.[2] Our motivation for this data was that (1) reviews of scientific articles should contain a strictly factual and precise argumentation and (2) this register has not yet been tackled in argumentation mining, to the best of our knowledge.

Amazon Camera Reviews is a collection of 16 product reviews from Amazon written by users. While reviews have been heavily exploited in the sentiment analysis field, there have been only few works on argumentation mining in this genre.

Twitter is a collection of 50+50 random Tweets related to 'Brexit' collected over the period of two weeks in April 2016 using hashtags such as `#strongerin` or `#leaveeu`. The collection was cleaned in terms of removing all re-tweets, tweets containing images or links to articles, and keeping only tweets without mentions of other Twitter accounts. This filtering step should have ensured that the tweets were written by the users 'from scratch' with the intention to support either staying in the EU or leaving it.

Policy Making is a collection of material derived from policy-making consultations that are carried out by the European Union on the topic of copyright for academic access. So far as we know, there have been no works on argumentation mining in this genre. The outcomes of the consultations are used to develop directives. The dataset is made up of responses by stakeholders to particular policy-making queries circulated by the commission. Thus, different stakeholders, e.g. companies and academic libraries, may have very different and contrastive views. The participants were provided with the *Green Paper: Copyright in the Knowledge Economy*, three sample responses, and two questions to focus attention on (questions 4 and 5).[3]

Grab Bag

Along with the four data sources presented above, we provided a 'grab bag' with a more diverse collection of data which were either solicited from the seminar participants or provided as additional datasets by the organizers. The following list introduces data that were discussed by at least one group in the break-out sessions.

Debate Portals is a subset of data (variant A) provided for the Unshared Task for the 3rd Workshop on Argument Mining co-located with ACL 2016 in Berlin.[4] Samples originate from a two-sided debate portal createdebate.com.

Opening and closing speeches from Oxford-style debates is a subset of data (variant B) provided for the Unshared Task for the 3rd Workshop on Argument Mining co-located with ACL 2016.

News Editorials with two articles about the currency in the Scottish Independence Referendum.

Persuasive Essay Revisions are two text files which represent the first and second drafts of a persuasive essay (related to educationally oriented argument mining for writing support).[5]

---

[2] http://papers.nips.cc/
[3] On http://ec.europa.eu/internal_market/copyright/copyright-infso/index_en.htm#maincontentSec2 see *Green Paper* and *The Replies to the Public Consultation* for ENPA, FAEP, and British Library.
[4] https://github.com/UKPLab/argmin2016-unshared-task/
[5] Contributed by Diane Litman

Medial Article is an excerpt from a scientific article containing an abstract, author summary, introduction, and discussion.[6]

Observations and discussion

This section attempts to summarize the main observations and possible future tasks for each particular dataset collected across groups and rings. During the final plenary discussion, we performed a poll in order to assess each data type with respect to following critical questions.

Critical Questions

Appropriateness Is the dataset "argumentative" enough to be worth investigating further?
Impact Is the task "attractive" enough to gain more visibility for our community?
Novelty Has the task already been tackled in computational argumentation or Computational Linguistics?
Visibility Would a shared task on this data attract researchers also from outside our community?
Comprehensibility What amount of domain knowledge is required to understand the data?
Reproducibility Do we have enough resources to be annotated and made freely available?
Scalability Is the task feasible for annotation by crowds?
Relativity Is the representation relative to individuals or groups?
Computability Can the representation/annotation/analysis be formalized?
Community Relevance Is the corpus and its analysis relevant to audience / community X?

Scientific Paper Reviews

It was pointed out by many that scientific reviews and author responses do follow some kind of latent structure. One particular direction of analysis thus can focus on identifying structural parts that deal with *clarity*, *originality*, and similar aspects of a scientific piece of work.

Proposed downstream tasks can tackle the primary outcome of a review, namely recognizing whether the submitted paper is good or bad; this coarse grained analysis, however, partly overlaps with sentiment analysis reviews. For that, the final marks of a review can be used as a proxy for annotated data. A more fine-grained task that goes deeper into the argumentation structures might thus involve *inconsistencies* between the the reviewer arguments (comments) and the final related marks given. Moreover, identifying reasons for *rejection*, *acceptance*, or *revisions* were identified as meaningful application outcomes.

Taking into account the dialog between the author's response and the original reviews, there are two options for analysis. The first use case is to analyze authors' rebuttals to the reviewers' comments in order to reveal which arguments were addressed. The second use case is to compare reviews from several reviewers and find contradictions or inconsistencies.

Overall, the intention in argumentative analysis of reviews (and optionally authors' responses) is to support scientific reviewing. The plenary discussion tackled not only the possible argument mining tasks and applications in scientific reviews, but also incorporating more elaborated reviewing strategies (such as providing reviewers with argument templates, or even some sort of claim and premise form). However, these issues tackle fundamental

[6] Contributed by Nancy Green

reviewing policies in the particular communities and are beyond the scope of the discussion here.

The plenum agreed that this dataset is very *appropriate* for further studying and a possible domain for a shared task, and is *novel* with a possible big *impact*. It would also help gain more *visibility* of computational argumentation. However, the level of expertise required for annotating the data is high, which might hinder *scalability*.

Amazon Camera Reviews

While product reviews have been a widespread data source for example in the sentiment analysis field, the major concern of the task participants was whether this genre is relevant to computational argumentation at all. Since fine-grained sentiment analysis (e.g., aspect-based sentiment) can already deal with a good level of information extraction from reviews and has reached maturity to be deployed in business, the added value of analyzing arguments in reviews remains an open question. Therefore the obvious tasks, such as extracting reasons for buyer's decisions seem not to be attractive enough, as the reasoning in reviews is usually based on listing pros and cons. Usefulness of reviews and its relation to argumentation was discussed during the plenary session but no clear consensus was reached.

Twitter

The main message taken from analyzing the Brexit-related tweets was that Twitter users try to by as cryptic as possible and rather than arguing, they tend to show off. One big obstacle is the need of background knowledge for the given topic in order to detect the stance or assess relevance of the presented arguments. Only occasionally, full-fledged arguments were presented. However, some potential tasks using Twitter data might include stance detection (already introduced at SemEval-2016), detecting a scheme of an argument, or mining relevant controversial sub-topics (such as the degree of control, costs of campaigns, etc., in case of Brexit). Overall, Twitter has not been found to be strongly appropriate as a shared task for argument mining in the near future, however, it's specifics (short messages, language, instant nature) make it a very challenging resource.

Policy making

While the argumentative and legal nature of the consultation make this dataset an interesting resource for studying argumentation, the task participants spent most of the effort on understanding the context, argumentative intention and decoding the main message. It was observed that the arguments are often about values as are other policy-making procedures. The nature of this dataset was found to be too difficult for the shared task, as it requires lots of background information to make sense of the subtle lines of argumentation. Therefore a possible high-level task could involve intention recognition. Overall, the dataset was found to be very appropriate for future argumentation mining research, if broken down into smaller feasible sub-tasks.

Future Work

As previous section highlighted, the discussion about the unshared untask highlighted criteria to identify relevant corpora rather than analysis tasks, though the identification of the corpora was largely determined by efforts to carry out some analysis. Thus, it remains for future work to:

- identify a range of tasks to exercise on a given corpus.
- develop ways to distinguish argumentative from non-argumentative texts as well as different genres within corpora of argumentative texts.
- offer evaluation metrics for multi-dimensional aspects of argumentation mining.
- incorporate context, point-of-view, background, and implicit/enthymatic information into argumentation mining.

References

1. Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado, 2015. Association for Computational Linguistics.
2. Noah A Smith, Claire Cardie, Anne Washington, and John Wilkerson. Overview of the 2014 NLP Unshared Task in PoliInformatics. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 5–7, Baltimore, MD, USA, 2014. Association for Computational Linguistics.
3. Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. Overview of the DSL Shared Task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria, 2015. Association for Computational Linguistics.

Participants

- Laura Alonso Alemany
National Univ. – Córdoba, AR
- Michal Araszkiewicz
Univ. Jagiellonski – Krakow, PL
- Kevin D. Ashley
University of Pittsburgh, US
- Pietro Baroni
University of Brescia, IT
- Floris Bex
Utrecht University, NL
- Katarzyna Budzynska
Polish Academy of Sciences – Warsaw, PL
- Elena Cabrio
Laboratoire I3S – Sophia Antipolis, FR
- Claire Cardie
Cornell University, US
- Federico Cerutti
Cardiff University, GB
- Giorgos Flouris
FORTH – Heraklion, GR
- Nancy L. Green
University of North Carolina – Greensboro, US
- Iryna Gurevych
TU Darmstadt, DE
- Ivan Habernal
TU Darmstadt, DE
- Graeme Hirst
University of Toronto, CA
- Eduard H. Hovy
Carnegie Mellon University – Pittsburgh, US
- Anthony Hunter
University College London, GB
- Diane J. Litman
University of Pittsburgh, US
- Bernardo Magnini
Bruno Kessler Foundation – Trento, IT
- Robert Mercer
University of Western Ontario – London, CA
- Marie-Francine Moens
KU Leuven, BE
- Smaranda Muresan
Columbia Univ. – New York, US
- Vincent Ng
University of Texas at Dallas, US
- Sebastian Padó
Universität Stuttgart, DE
- Fabio Paglieri
CNR – Rome, IT
- Alexis M. Palmer
Universität Heidelberg, DE
- Andreas Peldszus
Universität Potsdam, DE
- Ariel Rosenfeld
Bar-Ilan University – Ramat Gan, IL
- Patrick Saint-Dizier
CNRS-Paul Sabatier University – Toulouse, FR
- Jodi Schneider
University of Pittsburgh, US
- Edwin Simpson
TU Darmstadt, DE
- Noam Slonim
IBM – Haifa, IL
- Jan Šnajder
University of Zagreb, HR
- Manfred Stede
Universität Potsdam, DE
- Benno Stein
Bauhaus-Universität Weimar, DE
- Simone Teufel
University of Cambridge, GB
- Francesca Toni
Imperial College London, GB
- Leon van der Torre
University of Luxembourg, LU
- Serena Villata
Laboratoire I3S – Sophia Antipolis, FR
- Vern R. Walker
Hofstra Univ. – Hempstead, US
- Zhe Yu
University of Luxembourg, LU