Report of Dagstuhl Seminar 15512

# Debating Technologies

**Edited by**

**Iryna Gurevych[1], Eduard H. Hovy[2], Noam Slonim[3], and Benno Stein[4]**

1 **TU Darmstadt, DE,** gurevych@ukp.informatik.tu-darmstadt.de
2 **Carnegie Mellon University – Pittsburgh, US,** hovy@cmu.edu
3 **IBM – Haifa, IL,** noams@il.ibm.com
4 **Bauhaus-Universität Weimar, DE,** benno.stein@uni-weimar.de

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15512 "Debating Technologies". The seminar brought together leading researchers from computational linguistics, information retrieval, semantic web, and database communities to discuss the possibilities, implications, and necessary actions for the establishment of a new interdisciplinary research community around debating technologies. 31 participants from 22 different institutions took part in 16 sessions that included 34 talks, 13 themed discussions, three system demonstrations, and a hands-on "unshared" task.

**Seminar** December 13–18, 2015 – http://www.dagstuhl.de/15512
**1998 ACM Subject Classification** I.2 Artificial Intelligence, I.2.7 Natural Language Processing, H.3.3 Information Search and Retrieval, I.2.9 Robotics
**Keywords and phrases** Computational Argumentation, Discourse and Dialogue, Debating Systems, Human-machine Interaction, Interactive Systems
**Digital Object Identifier** 10.4230/DagRep.5.12.18
**Edited in cooperation with** Khalid Al-Khatib

## 1 Executive Summary

*Iryna Gurevych*
*Eduard H. Hovy*
*Noam Slonim*
*Benno Stein*

Why do people in all societies argue, discuss, and debate? Apparently, we do so not only to convince others of our own opinions, but because we want to explore the differences between our own understanding and the conceptualizations of others, and learn from them. Being one of the primary intellectual activities of the human mind, debating naturally involves a wide range of conceptual capabilities and activities, ones that have only in part been studied from a computational perspective in fields like computational linguistics and natural language processing. As a result, computational technologies supporting human debating are scarce, and typically still in their infancy. Recent decades, however, have seen the emergence and flourishing of many related and requisite computational tasks, including sentiment analysis, opinion and argumentation mining, natural language generation, text summarization, dialogue systems, recommendation systems, question answering, emotion recognition/generation, automated reasoning, and expressive text to speech.

This Dagstuhl seminar was the first of its kind. It laid the groundwork for a new interdisciplinary research community centered around debating technologies – computational technologies developed directly to enhance, support, and engage with human debating. The seminar brought together leading researchers from relevant communities to discuss the future of debating technologies in a holistic manner.

The seminar was held between 13 and 18 December 2015, with 31 participants from 22 different institutions. The event's sixteen sessions included 34 talks, thirteen themed discussions, three system demonstrations, and a hands-on "unshared" task. Besides the plenary presentations and discussions, the program included several break-out sessions and mock debates with smaller working groups. The presentations addressed a variety of topics, from high-level overviews of rhetoric, argument structure, and argument mining to low-level treatments of specific issues in textual entailment, argumentation analysis, and debating-oriented information retrieval. Collective discussions were arranged for most of these topics, as well as on more forward-thinking themes, such as the potential and limitations of debating technologies, identification of further relevant research communities, and plans for a future interdisciplinary research agenda.

A significant result of the seminar was the decision to use the term computational argumentation to put the community's various perspectives (argument mining, argument generation, debating technologies, etc.) under the same umbrella. By analogy with "computational linguistics", "computational argumentation" denotes the application of computational methods for analyzing and synthesizing argumentation and human debate. We identified a number of key research questions in computational argumentation, namely:

- How important are semantics and reasoning for real-world argumentation?
- To what extent should computational argumentation concern itself with the three classical rhetorical appeals of ethos (appeal to authority), pathos (appeal to emotion), and logos (appeal to reason)? Is it sufficient to deal with logos, or is there some benefit in studying or modelling ethos and pathos as well?
- What are the best ways of dealing with implicit knowledge?

A number of discussion questions at the seminar followed from these points, particularly in relation to the data and knowledge sources required for implementing and evaluating computational argumentation systems. For example, are currently available datasets sufficient for large-scale processing or for cross-language and cross-domain adaptation? Can we reliably annotate logos, ethos, and pathos? In any case, what sort of data would be considered "good" for a shared task in computational argumentation? Is it possible for computational argumentation to repeat the recent successes of "deep" natural language processing by employing shallow methods on large masses of data? How does cultural background impact human argumentation, and is this something that computational models need to account for? Finding the answers to these and other questions is now on the agenda for our burgeoning research community.

## 2 Table of Contents

**Executive Summary**

**Overview of Talks**

**Working groups**

**Panel discussions**

# 3 Overview of Talks

## 3.1 The Web as a Corpus of Argumentation

*Khalid Al-Khatib (Bauhaus-Universität Weimar, DE)*

Computational argumentation approaches are usually trained and evaluated on manually annotated texts. However, manual annotation for argumentation is particularly intricate and expensive, and practically infeasible for large numbers of texts and for the existing diversity of domains. As an alternative to manual annotation, we consider four types of web resources: social networks (Reddit, Facebook), discussion forums (Idebate), wiki (Wikipedia), and text extracted from web crawls. We discuss how to exploit them for automatic acquisition of labeled texts to address three computational argumentation tasks: the identification of controversial topics, of argument units and relations, and of positive and negative consequences.

### References

1 Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. *Cross-Domain Mining of Argumentative Text through Distant Supervision*. In 15th Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL'16) (to appear). Association for Computational Linguistics, San Diego, CA, USA, 2016.

## 3.2 Evidence Detection

*Carlos Alzate (IBM Research – Dublin, IE)*

**Joint work of** Ehud Aharoni; Carlos Alzate; Lena Dankin; Mitesh M. Khapra; Ruty Rinott; Noam Slonim
**Main reference** R. Rinott, L. Dankin, C. Alzate, M. M. Khapra, E. Aharoni, and N. Slonim, "Show Me Your Evidence – an Automatic Method for Context Dependent Evidence Detection," in Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing (EMNLP'15), pp. 440–450, Association for Computational Linguistics, 2015.
**URL** http://aclweb.org/anthology/D/D15/D15-1050.pdf

A methodology to automatically detect evidence (also known as premise) supporting a given claim in unstructured text is presented. This task has many practical applications in persuasion enhancement and decision support in a variety of domains. First, an extensive benchmark dataset specifically tailored for this task is introduced. Then, we propose a system architecture based on supervised learning to address the evidence detection task. Experimental results are promising and show the applicability of the proposed scheme.

### References

1 Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. *Show Me Your Evidence – an Automatic Method for Context Dependent Evidence Detection*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015.

2 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. Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 64–68, 2014.

## 3.3 Analogies as a Base for Knowledge Exchange and Argumentation

*Wolf-Tilo Balke (TU Braunschweig, DE)*

**Joint work of** Wolf-Tilo Balke; Silviu Homoceanu
**Main reference** S. Homoceanu and W.-T. Balke, "A Chip Off the Old Block – Extracting Typical Attributes for Entities based on Family Resemblance," In Proc. of the 20th Int'l Conf. on Database Systems for Advanced Applications (DASFAA'15), LNCS, Vol. 9049, pp. 493–509, Springer, 2015.
**URL** http://dx.doi.org/10.1007/978-3-319-18120-2_29

Analogies and analogical reasoning are major tools for the efficient communication between humans, and in particular for knowledge exchange and argumentation. In a simplistic definition the finding, exchanging, and understanding of analogies refers to a complex cognitive process of transferring information or meaning from one particular subject (the source) to another particular subject (the target). In the case of knowledge exchange, this means that without actually possessing deeper knowledge about a target entity or concept, the correct decoding of analogies allows to transfer some specific characteristics, attributes, or attribute values from a given source well known to all participants of the exchange; this effect is especially helpful in interdisciplinary discourses. In argumentation the benefit of analogies mostly lies in reducing complexity, for example when simplifying things or focusing a discussion by leaving out unnecessary details or when using analogies in the sense of precedence and arguing for similar measures to be taken in similar cases.

In a first phase of our work we are restricting analogies to information about entities, often referred to as entity summaries and provided in structured form, for instance by schema.org or Google's knowledge graph. However, for the later use in analogies not all properties of some entity can be used, since on one hand the intended property or concept has to be transferable over several cases of entities of the same category, and on the other hand it has to be widely known such that the analogy can be easily understood by the intended audience. To this aim we discuss how to derive a common entity structure or schema comprising attributes typical for entities of the same or similar entity type. To find out what is really typical, the definition of a practical measure for attribute typicality is needed (e.g., the measure derived from cognitive psychology presented in [1]. Since there is a wide variety of entity types and a manual inspection and classification might prove too expensive, further questions to be solved are the basic extraction of analogies from text and – where applicable – the generalization of analogies to other entities of a kind with the intention of finding out exactly which attributes or entity characteristics are essential for a certain the analogy to work.

### References

1 Silviu Homoceanu and Wolf-Tilo Balke. *A Chip Off the Old Block – Extracting Typical Attributes for Entities based on Family Resemblance*. In Proc. of the 20th International Conference on Database Systems for Advanced Applications (DASFAA), Hanoi, Vietnam, 2015.

## 3.4 Claim Generation

*Yonatan Bilu (IBM – Haifa, IL) and Noam Slonim (IBM – Haifa, IL)*

Computational Argumentation has two main goals – the detection and analysis of arguments on the one hand, and the synthesis or generation of arguments on the other. Much attention has been given to the former – mostly under the title of argumentation mining, but considerably less to the latter. Several models have been suggested for the structure of an argument – dating back to Aristotle, and in modern times the Toulmin model, or the more detailed Argumentation Schemes. A key component in all these models is the Conclusion or Claim of the argument. Thus, a key component in synthesizing arguments is the synthesis of claims.

One way to obtain claims for the purpose of generating arguments is by employing argumentation mining to detect claims within an appropriate corpus. While in specific cases, such as in the legal domain, one can use a corpus which is argumentative by nature and contains many claims, claim detection in the general case appears to be a hard problem. Thus, it is interesting to explore if – for the sake of synthesis – there may be other ways to generate claims.

Here we suggest such a method. We first go over a set of simple labeled claims for numerous topics (relatively short claims with exactly one verb), and extract the predicate part of these sentences. We call this the Predicate Lexicon. Given a new topic, we synthesize claims using a two step algorithm: First we construct candidate claims by constructing sentences whose subject is the new topic and the predicate is one from the Predicate Lexicon which bears some semantic similarity to the new topic. Second, we use a logistic regression classifier to determine whether the claim candidate is a coherent claim and appropriate for the topic at hand. The classifier is trained on candidate claims from the generation phase which were labeled via Amazon's Mechanical Turk. While these annotations are rather noisy, we are able to distill from them a relatively consistent labeling, for which we obtain surprisingly good results.

## 3.5 Emotions in Argumentation

*Elena Cabrio (INRIA Sophia Antipolis – Méditerranée, FR)*

Argumentation is a mechanism to support different forms of reasoning such as decision making and persuasion and always cast under the light of critical thinking. In the latest years, several computational approaches to argumentation have been proposed to detect conflicting information, take the best decision with respect to the available knowledge, and update our own beliefs when new information arrives. The common point of all these approaches is that they assume a purely rational behavior of the involved actors, be them humans or artificial agents. However, this is not the case as humans are proved to behave differently, mixing rational and emotional attitudes to guide their actions. Some works have claimed that there exists a strong connection between the argumentation process and the emotions felt by people involved in such process. We advocate a complementary, descriptive and experimental method, based on the collection of emotional data about the way human reasoners handle emotions during debate interactions. Across different debates, people's argumentation in plain English is correlated with the emotions automatically detected from the participants, their engagement in the debate, and the mental workload required to debate. Results show several correlations among emotions, engagement and mental workload with respect to the argumentation elements. For instance, when two opposite opinions are conflicting, this is reflected in a negative way on the debaters' emotions. Beside their theoretical value for validating and inspiring computational argumentation theory, these results have applied value for developing artificial agents meant to argue with human users or to assist users in the management of debates.

### References

1 Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon: *Emotions in Argumentation: an Empirical Evaluation*. IJCAI 2015. Buenos Aires, Argentina, pp. 156–163, 2015.

## 3.6 Profiling for Argumentation

*Walter Daelemans (University of Antwerp, BE)*

**Joint work of** Walter Daelemans; Frederik Vaassen
**Main reference** F. Vaassen and W. Daelemans, "Automatic Emotion Classification for Interpersonal Communication," in Proc. of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA'11), pp. 104–110, Association for Computational Linguistics, 2011.
**URL** http://www.aclweb.org/anthology/W11-1713

In order to adapt argumentation to the intended audience, we have to detect emotional states and personality aspects of this audience, preferably multimodal (e.g. in robot – person interaction) but in the simplest case only from text. As an early example of this, I describe the deLearyous project in which the argumentation system has to detect the communicative stance of the dialogue partner (above vs. below, together vs. against) from text only, according to a particular communication model (Leary's Rose). Although this turns out to be possible above chance level and at an accuracy higher than annotation agreement in people, the results were not good enough to be useful in practical systems. After a brief overview of the state of the art and main problems in computational stylometry and the need for balanced corpora in this field, I describe a more promising multimodal approach with NAO robots that is being started up at CLiPS. At this stage, the robot can adapt a more introverted or extraverted interaction style both in posture and language generation.

### References

1 Walter Daelemans. *Explanation in Computational Stylometry*. In: Gelbukh Alexander (Ed.): Computational Linguistics and Intelligent Text Processing, 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II. Springer 2013 Lecture Notes in Computer Science ISBN 978-3-642-37255-1, pp. 451–462, 2013.

## 3.7 Debating-Oriented Information Retrieval (Towards the WHY Search Engine)

*Norbert Fuhr (Universität Duisburg-Essen, DE)*

In this talk, we discuss the need for more advanced search engines that are able to answer why questions. Major applications would be all kinds of decision support, as well as helping in understanding and learning. Current Web search engines are mainly word-based, thus they can give satisfying answers to why queries only when there are single Web pages containing a comprehensive answer. In any case, however, they prefer positive answers over negative ones, and they suffer from a click-through bias. For developing WHY search engines, a number of research issues have to be addressed, such as argument representations suitable for information retrieval, methods for de-duplication of arguments as well as for estimating the importance and credibility of arguments, personalized and interactive retrieval of arguments.

## 3.8 Basic Concepts of Argumentation

*Graeme Hirst (University of Toronto, CA)*

This talk reviews some of the basic concepts of argumentation, and establishes some terminology. Unfortunately, there is much inconsistency and fuzziness in the literature with regard to terminology, and many terms are vague or used somewhat differently by different people. In assembling this set of basic concepts, I've drawn primarily from [1] and [3]. Points covered are the following:

- The distinction between individual *arguments* and *argumentation* as a sequence of moves, and how argumentation differs from the use of formal logic.
- The elements of arguments: assertions that may be *claims* or *reasons*, and which may be explicit or left implicit – that is, the argument may be an *enthymeme.*
- The four basic kinds of argument structure – *convergent, linked, serial,* and *divergent* – and diagrammatic notations for them.
- Three basic types of argument: *deductive*, probabilistically *inductive*, and *presumptive.*
- The notion of *strategic maneuvering* to achieve both dialectical and rhetorical goals in an argument: putting forward claims and arguments, asking questions, casting doubts, attacking opposing arguments.
- Three different kinds of attack: *rebuttals, undercutters*, and *defeaters.*
- The notions of *relevance, rationality, commitment* and *burden of proof* in argumentation.
- The embodiment of these ideas in the 10 rules of critical discussion proposed in the *Pragma-Dialectical Theory of Argumentation* [2].
- The *Toulmin model* of argumentation, with three kinds or levels of reasons in support of a claim – *grounds* or *data*, *warrant*, and *backing* – which may have *modal qualifiers* and *rebuttals.*
- *Argumentation schemes* as templates for common forms of argument, mostly presumptive or defeasible forms. Argumentation schemes are explicated at greater length in my second talk.

### References

1 Douglas Walton. *Fundamentals of Critical Argumentation*. Cambridge University Press. 2006.

2 Frans van Eemeren, Rob Grootendorst, and Snoeck Henkemans. *Argumentation: Analysis, Evaluation, Presentation*. Lawrence Erlbaum Associates, 2006.

3 Stephen Toulmin, Richard Rieke, and Allan Janik. *An Introduction to Reasoning*. Macmillan, 1979.

## 3.9 Introduction to Argumentation Schemes

*Graeme Hirst (University of Toronto, CA)*

This talk reviews some of the basic concepts of argumentation schemes and recent research in NLP on recognizing them. Points covered are the following:

- Argumentation schemes are templates for common forms of argument; they are mostly presumptive or defeasible, and may even be "fallacious". Examples include ad hominem arguments, argument from generic division, and argument from expert opinion.
- Argumentation schemes reflect real-world argumentation, and reject the hegemony of formal logicist approaches (which are included, but are not dominant).
- Although argumentation schemes go back to Aristotle, recent conceptions are due to [3] and [4]. The latter catalogue 65 schemes.
- Each scheme in [4] catalogue is given a set of *critical questions*, which can be used as challenges to premises of arguments in the scheme and as suggestions for missing premises of enthymemes.
- [1] developed a system that recognized five common argumentation schemes. Their method assumed that the argumentative text and its premise and conclusion have already been identified. The features they used were certain surface characteristics (used for all five schemes), and scheme-specific features such as keywords, textual similarity, and lexical sentiment. They achieved medium to high accuracy for discriminating most of the schemes from the others.
- [2] removed the assumption that premises and conclusion have been previously identified, using a cascade of weak methods to identify both the components and the argumentation scheme. The methods included discourse connections and features similar to those of [1]. They evaluated on two schemes with medium to high accuracy.

### References

1 Vanessa Wei Feng and Graeme Hirst. *Classifying Arguments by Scheme*. Proceedings of 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, pp. 987–996, 2011.

2 John Lawrence and Chris Reed. *Combining Argument Mining Techniques*. Proceedings of the 2nd Workshop on Argumentation Mining, Denver, Colorado, pp. 127–136, 2015.

3 Chaim Perelman and Lucie Olbrechts-Tyteca. *The New Rhetoric*. University of Notre Dame Press, 1969.

4 Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

## 3.10 Expertise and Argumentative Personas: Detection and Generation

*Eduard H. Hovy (Carnegie Mellon University – Pittsburgh, US)*

**Joint work of** Eduard H. Hovy; Siddharth Jain

In contentious discussions, in parallel with their formal stance, the partners provide unconscious signals about their expertise, commitment, cooperativeness, etc. In this talk we describe our work on the automated detection of different 'argumentation personas', including the level of Leadership, Rebelliousness, Sense-making, etc. To estimate Leadership, we develop two different models, one counting the degree to which participants adopt the terminology and the stance of a person throughout the course of the arguments, and the other counting the number of times a person's contribution brings an ongoing disagreement to an end. To estimate Sense-making, our model integrates several factors, including reference to rules of argumentation, presence or absence of ad hominem attacks, adherence to the topic, etc. We test these ideas on two social media arenas: Wikipedia's Articles for Deletion discussions and 4forums.com political and religious arguments.

## 3.11 Opinions and Why they Differ from Sentiment

*Eduard H. Hovy (Carnegie Mellon University – Pittsburgh, US)*

NLP has witnessed a rapidly growing desire for automated sentiment/opinion detection systems. But the absence of clear definitions of the major topics under discussion hampers serious work. In this brief talk I provide definitions for two types of Opinion – Judgment (with values like good, bad, mixed, neutral; typically called sentiment) and Certainty (true, false, possible, etc.; typically called epistemic judgments). Any argumentation or debating system needs to be able to handle both kinds of Opinion appropriately. I describe their parallel structure (Holder, Topic, Claim, Valence) and the still unaddressed facet of Reason.

## 3.12 What is Argumentation and Rhetoric?

*Eduard H. Hovy (Carnegie Mellon University – Pittsburgh, US)*

Too often in the past, the NLP community has opened up a new topic without due regard to its general nature. Typically the community focuses on some specific easy-to-computationalize aspects without regard to the broader context in which their work exists. The resulting narrow focus often limits the scope and applicability of their work. This workshop focuses on debating, which is a specific type of argumentation. The novelty of this topic presents the danger of rapid narrowing and overspecialization. Therefore, in this introductory talk I provide a general background on Argumentation. Starting with what argumentation and rhetoric are (as defined by Aristotle and expended through the ages), we consider different types/families of argumentation (from antiquity to now, including subtypes like legal arguments, academic discussions, and debates). Each argument has the same basic structure, consisting of premises and interpretive values for them, plus relations that express how one premise supports or attacks another. But each type of argument has evolved its own internal logic and stereotypical structure, and debating systems need to be able to recognize when one or another of these modes is adopted by an adversary. This then leads naturally to the principal theoretical questions to be addressed by argumentation systems, and by debating systems in particular, including: how one discovers argumentative text in the wild (argument retrieval), how one automatically analyzes such examples ("argumentation mining"), how one maintains and updates the argument structure as it evolves, how one automatically generates premises in arguments ("argument generation"), and how one can employ a sub-case of argumentation for practical purposes ("debating technologies").

## 3.13 What is Debating Technologies

*Noam Slonim (IBM – Haifa, IL)*

Why do people in all societies argue, discuss, and debate? Apparently, we do so not only to convince others of our own opinions, but because we want to explore the differences between our own understanding and the conceptualizations of others, and learn from them. Being one of the primary intellectual activities of the human mind, debating therefore naturally involves a wide range of conceptual capabilities and activities, ones that have only in part been studied from a computational perspective.

In this talk I described recent work done by IBM Research to develop Debating Technologies, defined as computational technologies developed directly to enhance, support, and engage with human debating. We further discussed the inter-relations of Debating Technologies and Argumentation Mining, and their role in the more general emerging field of Computational Argumentation.

## 3.14 Communication of Debate Aspects to Different Audiences

*Brian Plüss (The Open University – Milton Keynes, GB)*

The Election Debate Visualisation (EDV) project (http://edv-project.net/) aims to improve democratic citizenship, making televised election debates more accessible and engaging by giving viewers tools to make sense of complex political argumentation. The project brings together research from political communication, computational linguistics, collective intelligence and design in order to provide enhanced, interactive online debate replays [1] . Centered around the citizens' democratic expectations about election debates [2], data from several sources related to a televised debate (video, transcript, live audience responses, tweets, etc.) are analyzed. The results of these analyses are then shown as interactive visualisations, in synchrony with the video of the debate.

This talk focuses on the challenges in producing visualisations that are suitable for a wide range of audiences, from domain experts and data scientists, to politics students and the general public. The issues and proposed solutions are illustrated with a demo of the tools developed by the EDV project, covering debate aspects such as: computer supported argument visualisation [3], debate rule compliance and fair play [5][6], and a novel method for capturing real-time audience feedback to media events [7].

**References**

1 Brian Plüss and Anna De Liddo. *Engaging Citizens with Televised Election Debates through Online Interactive Replays*. In Proceedings the ACM International Conference on Interactive Experiences for TV and Online Video, Brussels, Belgium, pp. 179–184, 2015.

2 Stephen Coleman and Giles Moss. *Rethinking Election Debates: What Citizens Are Entitled to Expect*. The International Journal of Press/Politics, 21(1), pp. 3–24, 2016.

3 Simon Buckingham Shum. *The Roots of Computer-Supported Argument Visualization*. Visualizing Argumentation. London: Springer-Verlag, pp. 3–24, 2003.

4 Anna De Liddo, Sándor Ágnes, and Simon Buckingham Shum. *Contested Collective Intelligence: Rationale, Technologies, and a Human-Machine Annotation Study*. Computer Supported Cooperative Work (CSCW), 21(4–5), pp. 417–448, 2012.

5 Brian Plüss, 2014. *A Computational Model of Non-Cooperation in Natural Language Dialogue*. Doctoral dissertation, The Open University, 2014.

6 Stephen Coleman, Simon Buckingham Shum, Anna De Liddo, Giles Moss, Brian Plüss, and Paul Wilson. *Rhetoric and the Rules of the Game*. EDV Project Briefing 2014.03. August 2014.

7 Stephen Coleman, Simon Buckingham Shum, Anna De Liddo, Giles Moss, Brian Plüss, and Paul Wilson. *A Novel Method for Capturing Instant, Nuanced Audience Feedback to Televised Election Debates*. EDV Project Briefing 2014.04. December 2014.

## 3.15 Argument(ation) and Social Context

*Vinodkumar Prabhakaran (Stanford University, US)*

Social context of an interaction often affects how its participants interact with one another. The social context may derive from a multitude of factors such as status, power, authority, experience, age and gender of the participants. Researchers in NLP have recently started looking into how these social factors affect various aspects of interactions such as politeness [2], dialog structure [3, 4], and level of commitment [1]. In this talk, I argue for the need to consider the effects of social context in argumentation, and what it means to the community of researchers working on computational argumentation. For example, [5] found important differences between the efficacy of argumentation patterns exhibited by men and women. They also suggest that these differences are not inherently tied to gender, but rather due to the power imbalances between genders. Another important aspect that will affect argumentation patterns is the prevalent culture within which it occurs. Beyond gaining potential sociolinguistics insights, research in this direction might also be of practical importance to argument mining systems. On the other hand, it is an interesting open question whether or not a debating technology should be agnostic to the social context while constructing an argument.

### References

1 Vinodkumar Prabhakaran. *Social Power in Interactions: Computational Analysis and Detection of Power Relations*. Columbia University, PhD thesis. 2015.

2 Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. *A cComputational Approach to Politeness with Application to Social Factors*. Proceedings of ACL, 2013.

3 Vinodkumar Prabhakaran and Owen Rambow. *Predicting Power Relations between Participants in Written Dialog from a Single Thread*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, 2014.

4 Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. *Gender and Power: How Gender and Gender Environment Affect Manifestations of Power*. Association for Computational Linguistics, 2014.

5 Patricia Hayes Bradley. *The Folk-linguistics of Women's Speech: An Empirical Examination*. Communication Monographs, pp. 73–90. vol 48, 1981.

## 3.16 The Role of Evidence in Debates

*Ruty Rinott (IBM – Haifa, IL)*

While the essentiality of claims in any argumentative text is well established, the need for evidence in such scenarios is less agreed upon. In this talk I reviewed current literature on the use of evidence in argumentative text, and specifically in debates. In general, there are many indications that using evidence in argumentative text enhances its persuasiveness. However, the degree of evidence influence, and the type of evidence to present depends on the speaker, the audience, and the topic under discussion. In practice, we observe that debaters who are topic experts tend to use evidence much more frequently than those who are not.

### References

1 R. A. Reynolds and J. L. Reynolds. *Evidence*. In The Persuasion Handbook: Developments in Theory and Practice, eds. J. P. Dillard, and M. Pfau, pp. 427–444, Thousand Oaks: Sage, 2002.

2 J. Hornikx. *A Review of Experimental Research on the Relative Persuasiveness of Anecdotal, Statistical, Causal, and Expert Evidence*. Studies in Communication Sciences 5:205–216, 2005.

3 Richard D. Rieke and Malcolm O. Sillars. *Argumentation and the Decision Making Process*. Harper Collins, 1984.

4 Z. Seech. *Writing Philosophy Papers*. Cengage Learning, 2008.

5 J. C. Reinard. *The Empirical Study of the Persuasive Effects of Evidence The Status After Fifty Years of Research*. Human Communication Research, 15:3–59. DOI: 10.1111/j.1468-2958.1988.tb00170.

6 P. H. Bradley. (1981). *The Folk-linguistics of Women's Speech: An Empirical Examination*. Communication Monographs, 48, pp. 73–90, 1981.

### 3.17 Detecting Argument Components and Structures

*Christian Stab (TU Darmstadt, DE) and Ivan Habernal (TU Darmstadt, DE)*

**Main reference** C. Stab and I. Gurevych, "Identifying Argumentative Discourse Structures in Persuasive Essays," in Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14), pp. 46–56, Association for Computational Linguistics, 2014.
**URL** http://www.aclweb.org/anthology/D14-1006

The detection of micro-level argument components and structures includes several subtasks which are independent of the particular text type or application scenario. First, the identification of argument components includes the recognition of text units which are relevant to the argumentation and the detection of its boundaries. Second, the identification of argument component types focuses on the identification of argumentative roles like e.g. claims, conclusions, premise or evidence. Finally, the identification of argumentation structures aims at identifying argumentative relations between argument components in order to detect the argumentative discourse structures in texts.

In this talk, we presented two different approaches for detecting argument components and structures. First, we introduce a corpus of web discourse annotated with an extended version of Toulmin's model of argument. The results of a semi-supervised approach using "argument space" features improve performance up to 90% in cross-domain and cross-register evaluation [1]. Second, we present the recent results of argument structure detection [2] in persuasive essays [3]. We show that joint modeling not only considerably improves the identification of argument component types and argumentative relations but also significantly outperforms a challenging heuristic baseline.

#### References

1 Ivan Habernal and Iryna Gurevych. *Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse.* In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), pp. 2127–2137, Lisbon, Portugal, 2015.

2 Christian Stab and Iryna Gurevych. *Annotating Argument Components and Relations in Persuasive Essays.* In: Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), pp. 1501–1510, Dublin, Ireland, 2014.

3 Christian Stab and Iryna Gurevych. *Identifying Argumentative Discourse Structures in Persuasive Essays.* In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 46–56, Doha, Qatar, 2014.

### 3.18 Existing Resources for Debating Technologies

*Christian Stab (TU Darmstadt, DE) and Ivan Habernal (TU Darmstadt, DE)*

Existing resources for Debating Technologies are predominantly present in the recent research field of Argumentation Mining. These corpora are usually tailored to a particular task, employ different annotation schemes, are limited to a particular text genre or exhibit different granularities of arguments or argument components. In this talk, we introduced a taxonomy for categorizing existing resources in order to facilitate the selection of existing benchmark resources and the definition of future annotation studies. In particular, our taxonomy structures existing resources by means of existing tasks in argumentation mining and the granularity of arguments (micro-level and macro-level) and argument components (clause-, sentence- and multi-sentence components). In addition, we provide an overview of several existing resources in order to identify requirements, challenges and visions for future resources.

## 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, pp. 64–68, Baltimore, MD, USA, 2014.

2 Elena Cabrio and Serena Villata. *NoDE: A Benchmark of Natural Language Arguments*. In Proceedings of COMMA 2014, pp. 449-450, 2014.

3 Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. *On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2236-2242, Lisbon, Portugal

4 Ivan Habernal and Iryna Gurevych. *Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2127–2137, Lisbon, Portugal, 2015.

5 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, pp. 1–11, Denver, CO, USA, 2015.

6 Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. 2007. *Identifying and Classifying Subjective Claims*. In Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines and Domains, pp. 76–81, Philadelphia, PA, USA.

7 Raquel Mochales-Palau and Marie-Francine Moens. *Argumentation Mining: The Detection, Classification and Structure of Arguments in Text*. In Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL'09, pp. 98–107, Barcelona, Spain, 2009.

8 Andreas Peldszus and Manfred Stede. *An Annotated Corpus of Argumentative Microtexts*. First European Conference on Argumentation: Argumentation and Reasoned Action, Portugal, Lisbon, 2015.

9 Chris Reed, Raquel Mochales-Palau, Glenn Rowe, and Marie-Francine Moens. *Language Resources for Studying Argument*. In Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC'08, pp. 2613–2618, Marrakech, Morocco, 2008.

10 Christian Stab and Iryna Gurevych. *Annotating Argument Components and Relations in Persuasive Essays*. In Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), pp. 1501–1510, Dublin, Ireland, 2014.

## 3.19 Discourse Structure and Argumentation Structure

*Manfred Stede (Universität Potsdam, DE)*

**Main reference** M. Stede, "Discourse Processing," Morgan and Claypool, 2011.

For finding arguments in natural language text, it is helpful to consider the relationship between the structure of argumentation and the more general notion of 'discourse structure', as it has been discussed in the CompLing community for a long time. A central goal of ascribing discourse structure to a text is in accounting for the text's coherence, i.e. it's "hanging together", which is often being regarded on three different levels of description:

- Referential continuity: A text keeps talking about the same things, i.e., the same discourse referents.
- Topic continuity: A text does not jump wildly between topics but addresses a discernible sequence of topics and their subtopics.
- Discourse relations: Adjacent sentences or spans of text tend to be in some semantic or pragmatic relation to each other.

In particular the third issue is related to argumentation. Discourse relations can be signalled at the text surface, most often by means of connectives: "Tom wants to buy a new apartment, *but* he is not rich enough." Connectives are conjunctions and various types of adverbials (e.g., 'however', 'therefore', 'afterwards'). In certain simple texts, identifying the connectives and their arguments (the text spans being related) is sufficient for deriving a complete discourse structure. Various studies have found, however, that most discourse relations are only implicit, i.e., there is no signal present. This often holds for temporal relations, as in "Tom ran to the station. He jumped onto the train." Similarly, causal relations do not need to be signalled when their presence can be easily inferred: "Tom's stomach was aching. He had eaten way too many French fries." In order to devise a theory of discourse structure around the notion of discourse relation, one has to answer at least these questions:

1. What is the set of relations?
2. How are the relations to be defined?
3. What predictions are being made on the structure of discourse?

Nowadays, the three most prominent approaches are the Penn Discourse Treebank (PDTB); Rhetorical Structure Theory (RST); and Segmented Discourse Representation Theory (SDRT).

The PDTB distinguishes four families of relations (temporal, contingency, comparison, expansion) and defines them mostly in terms of semantic descriptions that can be easily understood by annotators. This annotation mainly consists of identifying the presence of relations, marking the connective (if any) and the arguments. Since each relation is annotated individually, no claims on overall discourse structure are being made at this stage.

In contrast, RST posits that a tree structure results from recursively combining adjacent text spans with a discourse relation. There are two big families of relations: one for more pragmatic, intention-based ones; one for more semantic, content-based ones. An interesting claim of RST is that the vast majority of relations assign different weight to their two arguments: one is the 'nucleus' and most important for the writer's purposes; the other is the 'satellite' and has merely a supportive function.

SDRT has been developed from the perspective of formal semantics, and thus the relations are defined largely in terms of features of underlying event structure, etc. The analysis of a text yields a DAG, hence a discourse segment can play multiple roles, in contrast to RST.

Relations are split in two groups: ‘coordinating’ versus ‘subordinating’. They have different implications for the discourse structure, primarily for anaphoric accessibility.

For all three approaches, there exist automatic parsers. A discourse structure analysis can be a useful preparatory step for argumentation mining, in particular when the text has a certain minimal complexity, so that it makes sense to first break it down into portions and then map the discourse structure to a representation of the argumentation.

## 3.20 An Argument Relevance Model for IR

*Benno Stein (Bauhaus-Universität Weimar, DE)*

We report on research of how to develop a document ranking approach that explicitly models argument strength. We propose to combine state-of-the-art technologies for retrieval and mining to construct a special “argument graph” for a given query. This graph will be recursively evaluated, resembling ideas from the well-known algorithm PageRank, both for the combination of support and attack relations between multiple arguments, and for the assessment of “argument ground strength”.

Classical retrieval models provide the formal means of satisfying a user’s information need (typically a query) against a large document collection such as the web. These models can be seen as heuristics that operationalize Robertson’s probability ranking principle: “Given a query $q$, the ranking of documents according to their probabilities of being relevant to $q$ leads to the optimum retrieval performance.” The new generation of retrieval models that we envision goes into a more specific but possibly game-changing direction, supporting information needs of the following kind: “Given a hypothesis, what is the document that provides the strongest arguments to support or attack the hypothesis?”

Obviously, the implied kind of relevance judgments cannot be made based on the classical retrieval models, as these models do not capture argument structure. In fact, so far the question of how to exploit argument structure for retrieval purposes has hardly been raised, and we propose a comparably basic paradigm along with an operationalizable model that deal with the following aspects: canonical argument structures, interpretation functions, argument graphs, recursive relevance computation, and argument ground strength.

### References

1 Elena Cabrio and Serena Villata. *Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions.* 50th Annual Meeting of the Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, pp. 208–212, 2012.

2 Phan Minh Dung. *On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games.* Artificial Intelligence, vol. 77, pp. 321–357, 1995.

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

4 Benno Stein, Tim Gollub, and Maik Anderka. *Encyclopedia of Social Network Analysis and Mining, Retrieval Models.* Springer, pp. 1583–1586, 2014.

## 3.21 Paraphrasing

*Benno Stein (Bauhaus-Universität Weimar, DE)*

We consider the problem of automatically paraphrasing a text. To paraphrase means to rewrite the text's content whilst preserving the original meaning. From our point of view, handling paraphrasing is of major importance in the context of debating technologies.

While monological argumentation requires paraphrase recognition "only", dialogical argumentation requires both paraphrase recognition and generation. Approaches to tackle the former are word-level metrics, information retrieval metrics, or machine translation metrics, while approaches to latter include the learning from parallel corpora, combining translating and re-translating, text simplification and summarization, templating, and heuristic search in a so-called paraphrase-operator space. Obviously we are far away from solving the problems of paraphrase recognition and paraphrase generation in its entirety and may focus on "low-hanging fruits" first. Possible steps in this direction are to narrow the topic domain, to restrict to certain text genre, or to restrict to selected tasks, whereas an interesting task in regard with debating is to automate the dialog adaptation to the personality profile of the discussion partner.

### References

1 Nitin Madnani, J. Tetreault, and M. Chodorow. *Re-examining Machine Translation Metrics for Paraphrase.* Proceedings of the Conference of the North American Chapter of the ACL: Human Language Technologies, pp. 82–190, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012.

2 Benno Stein, Matthias Hagen, and Christof Bräutigam. *Generating Acrostics via Paraphrasing and Heuristic Search.* 25th International Conference on Computational Linguistics (COLING 14). Association for Computational Linguistics, Dublin, Ireland, pp. 2018–2029, 2014.

3 Asher Stern and Ido Dagan. *BIUTEE: A Modular Open-Source System for Recognizing Textual Entailment.* Proceedings of the ACL 2012 System Demonstrations, Association for Computational Linguistics, Jeju Island, Korea, pp. 73-78, 2012.

## 3.22 Enthymeme Reconstruction

*Simone Teufel (University of Cambridge, GB)*

**Joint work of** Olesya Razuvayevskaya; Simone Teufel

I presented initial work of my student Olesya Razuvayevskaya on linguistic enthymeme reconstruction. The reconstruction of enthymemes, in our definition of the task, is closely related to pragmatic effects. We report on the connection between presuppositions, conventional implicatures and omitted premises in mini-arguments. There is another obvious connection with entailment. Our observation is that in order to cut down the large search space of hypotheses, machine learning might identify contexts where the size of reasoning step is smaller than in the general case, for instance in contexts where the prepositional phrase "of course" is used, i.e., where the speaker explicitly announces the obviousness of a reasoning step. We reported on encouraging initial machine learning experiments using RTE-type features, where the (binary) task is to decide whether or not a “of course” segment corresponds to an enthymeme or not.

## 3.23 Analysis of Stance and Argumentation Quality

*Henning Wachsmuth (Bauhaus-Universität Weimar, DE)*

Argumentation mining is concerned with the *detection* of the argumentative structure of natural language texts in terms of the identification of argument units and the classification of relations between these units. In contrast, this talk considered the *analysis* of argumentative structure with the goal of using the structure to address specific argumentation-related tasks. Two such tasks (or families of tasks) are stance classification and the assessment of argumentation quality.

Stance classification aims to the determine the overall position of the author of a text towards a predefined topic. Mostly, only “for” and “against” are distinguished, sometimes also “none” or similar. Although assumed to be given, the topic is not necessarily mentioned in the text. Stance classification is connected to sentiment analysis among others, but a stance may also express sentiment on another topic–or none at all–and it depends on what the author argues to be true. Stance classification is important for debating technologies, especially because it is needed to identify pro and con arguments, although the restrictions to predefined topics and “for vs. against” scenarios might have to be revised in practice.

In the talk, the state of the art of classifying the stance of dialogical and monological argumentation was surveyed. For dialogical argumentation, existing approaches achieve an accuracy between 61% and 75% in the two-class scenario. Most of them analyze aspect-based or topic-directed sentiment to some extent. Some add knowledge about the stance of other texts of an author, whereas others exploit the dialog structure to identify opposing stances. Monological argumentation, on the other hand, puts more emphasis on the actual argumentative structure. For instance, Faulkner classifies 82% of all essay stances correctly based on a proprietary representation of arguments, derived from dependency parse trees. What is not captured so far, however, is the overall structure of an argumentation. Here, a model of the flow of rhetorical moves might be useful, which we found to be effective and robust in the related task of global sentiment analysis.

The assessment of argumentation quality is not a well-defined task yet. In overall terms, the argumentation quality of a text generally seems hardly measurable, because several quality dimensions may be important for arguments and argumentations, such as the logical correctness and completeness, the strength or convincingness, the comprehensibility, clarity, or similar. Some of these will be hard to assess in many real-world scenarios (e.g., logical correctness), and some will often depend on the preconceived opinion of the reader (e.g., strength). Still, there are different quality dimensions that have already been analyzed in previous research.

The talk aimed to give a first overview of research on the assessment of argumentation quality. Approaches have been proposed to determine which arguments are prominent or accepted. Other researchers study how deliberate a dialogical argumentation is, what evidence types can be found, or whether critical questions are answered in essays. For essays, in particular, argumentation-related dimensions have been investigated, e.g., thesis clarity and argument strength. Our recent research suggests that argumentation mining can be leveraged to better solve respective essay scoring tasks. In contrast, there are also very important aspects of argumentation quality that have hardly been approached in practice so far, such as the presence of fallacies or the impact of pathos and ethos as opposed to logos. And, finally, existing work on text quality should not be forgot when it comes to argumentative texts, e.g., regarding readability, text coherence, review deception, review helpfulness, or Wikipedia quality flaws.

## References

1 Filip Boltužić and Jan Šnajder. *Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity*. In Proceedings of the Second Workshop on Argumentation Mining, pp. 110-115, 2015.

2 Elena Cabrio and Serena Villata. *Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions*. In Proc. of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers – Volume 2, pp. 208–212, 2012.

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

4 Adam Robert Faulkner. *Automated Classification of Argument Stance in Student Essays: A Linguistically Motivated Approach with an Application for Supporting Argument Summarization*. Dissertation, City University of New York, 2014.

5 Valentin Gold, Mennatallah El-Assady, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. *Visual Linguistic Analysis of Political Discussions: Measuring Deliberative Quality*. Digital Scholarship in the Humanities, 2015.

6 Kazi Saidul Hasan and Vincent Ng. *Stance Classification of Ideological Debates: Data, Models, Features, and Constraints*. In Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP), pages pp. 1348–1356, 2013.

7 Isaac Persing, Alan Davis, and Vincent Ng. *Modeling Organization in Student Essays*. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 229–239, 2010.

8 Isaac Persing and Vincent Ng. *Modeling Thesis Clarity in Student Essays*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics – Volume 1: Long Papers, pp. 260–269, 2013.

9 Isaac Persing and Vincent Ng. *Modeling Prompt Adherence in Student Essays*. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics – Volume 1: Long Papers, pp. 1534–1543, 2014.

10 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, pp. 543–552, 2015.

11 Sarvesh Ranade, Rajeev Sangal, and Radhika Mamidi. *Stance Classification in Online Debates by Recognizing Users' Intentions*. In Proceedings of the SIGDIAL 2013 Conference, pp. 61–69, 2013.

12 Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. *Show Me Your Evidence – An Automatic Method for Context Dependent Evidence Detection*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 440–450, 2015.

13 Swapna Somasundaran and Janyce Wiebe. *Recognizing Stances in Online Debates*. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th

International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 – Volume 1, pp. 226–234, 2009.

14 Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. *Applying Argumentation Schemes for Essay Scoring*. In Proceedings of the First Workshop on Argumentation Mining, pp. 69–78, 2014.

15 Christopher W. Tindale. *Fallacies and Argument Appraisal. Critical Reasoning and Argumentation*. Cambridge University Press, 2007.

16 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, pp. 601–611, Lisbon, Portugal, 2015.

17 Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. *Modeling Review Argumentation for Robust Sentiment Analysis*. In Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers, pp. 553–564, 2014.

# 4 Working groups

## 4.1 Computational Argumentation Competitions and Data

*Khalid Al-Khatib (Bauhaus-Universität Weimar, DE) and Noam Slonim (IBM – Haifa, IL)*

**Joint work of** Khalid Al-Khatib; Elena Cabrio; Claire Cardie; Iryna Gurevych; Ivan Habernal; Chris Reed; Hinrich Schütze; Noam Slonim; Anita De Waard

This working group dealt with two issues in computational argumentation: the preparation of evaluations in terms of competitions as well as the creation of benchmark datasets for development and evaluation.

The envisioned competitions aim mainly at advancing research in the area of computational argumentation and at attracting new researchers from different communities to work in this area. However, conducting such a competition requires substantial effort; several requirements and challenges should be considered. First, the competition has to be clearly defined, attractive, and interesting. It should create buzz and it should preferably have a relatively low entry level. Second, the dataset used in the competition should enable a continuous evaluation of proposed systems before a final test dataset is published, and it should allow for clear and objective evaluation. And third, the competition's organizers should make sure to generate fair baseline results.

The working group also proposed some ideas for possible competitions, namely:

- The prediction of the level of persuasiveness of an argument(ation) regarding ethos, pathos, and logos in text, audio, or video.
- A competition on top of political debates with additional data coming from tweets.
- The prediction of what leads to a high citation rate of a claim in a scientific paper – in a time scale of 5–10 years.
- The detection of arguments (or claims, evidence, or similar) in a pre-specified set of documents for a given topic.
- The construction of arguments; given a controversial topic and a set of related documents, construct the most persuasive arguments.
- The prediction of the impact of a given argument over a particular audience – where an audience is characterized by nationality, age, gender, personality traits, or similar.

The working group suggested to have a call for challenges proposals as part of the ACL Argument Mining workshop 2016. The accepted proposals should be discussed as a poster session, and one or a few of them should turn into a real challenge during 2017.

Regarding the creation of benchmark datasets, the working group discussed three main points: First, the requirement that datasets proposed in the future should follow specific standards; being stable, having a DOI, and having a uniform format – even at the level of README files. Second, the importance of having a data repository for computational argumentation datasets. The existing "Argument Web" has been proposed for that purpose, considering that it may need to tighten up user control, keeping standards, using a uniform format, and so on. And third, the possibility of releasing datasets – under some reasonable standards – as a "requirement" for publication.

## 4.2 Debating Strategies

*Brian Plüss (The Open University – Milton Keynes, GB)*

**Joint work of** Jens Allwood; Carlos Alzate; Yonatan Bilu; Brian Plüss

In this session we discussed debating strategies. Understood as means to achieve individual goals in which someone wants to influence another person or group, debating strategies apply to sections of entire debates and can be thought as composed by building blocks: speech, dialogue acts, moves, etc. The suitability and efficacy of a strategy will depend on the activity type characterised, among others, by the initial situation, goal, cultural context and target audience.

We structured the discussion based on three types of high-level rhetorical goals a speaker might have: pathos, i.e. aiming that people behave in a certain way; ethos, i.e. establishing themselves as credible; and logos, i.e. establishing the reasonability or truth of a position. For pathos, strategies that could affect how people behave include: threatening (ad bacculum), luring or offering rewards, flattering, giving orders (assuming compliance is expected, that it's culturally acceptable, etc.), stating sympathy, exposing righteousness and morality (or lack thereof), appealing to solidarity or superiority, etc. For ethos, strategies that could affect how people think of the speaker include: stating personal experiences, stating expertise, presenting titles and credentials, getting somebody else to present the speaker's credentials, exposing good connections, establish mutual trust, etc. For logos, strategies that could make someone change what they believe to be the truth include: making claims, providing evidence, exposing implicit connections, establishing common ground and mutual understanding, expressing claims clearly, exposing righteousness or morality, etc.

## 4.3 Logic and Argumentation

*Simone Teufel (University of Cambridge, GB)*

**Joint work of** Walter Daelemans; Ido Dagan; Norbert Fuhr; Anette Frank; Christoph Lofi; Nils Reiter; Manfred Stede; Benno Stein; Simone Teufel

The topic of this break-out session was the connection between computational argumentation and logic and reasoning. We felt strongly that the connection is highly relevant: representation of the propositional contents of the argument, in the right form, and the ability to reason over such content, would allow for better manipulation of arguments both for analysis and synthesis, and thus advance research in computational argumentation. But of course the core question remains, how could enough world knowledge of the right kind be acquired, represented, and then made to work, reliably and robustly, in many interesting cases? Looking at this question in 2015/6, we also have to consider experiences from the past.

Most of the past research on reasoning was closely connected to using natural language as the source of the knowledge extracted. "Strong AI" approaches from the 70s and 80s include Winograd's SHRLDU, the Schankian approach to frames and stories and all its variants, which used rigid inference. What is attractive about these models is their depth and explanatory nature – when they work (namely in toy worlds). However, it proved impossible to scale them up to unlimited domains or larger settings. There were always too many knowledge gaps in the representation, which simply wasn't dense enough.

Later, the Cyc project tried to address the knowledge gap using shallow, large-scale knowledge extraction from large text corpora. As knowledge representation, however, it still used a symbolic language without grounding – one with atomic symbols that only "looked like" natural language, while in fact not being able to perform "linguistic" inference, for instance to systematically detect similarity with semantically close symbols, and so on. Natural languages in contrast are far too complex for us to model fully, as they can express ambiguity, vagueness, attribution, beliefs and many other "soft effects". It was mentioned in our group that information retrieval based on automatic indexing was the first time in history that large-scale knowledge representation was based on "real language", albeit in a very shallow form. Some kinds of knowledge probably cannot be meaningfully represented only in the form of keywords and simple connection strengths between them.

In the past, there have also been successes in the NLP field on "easy" tasks such as factoid question answering, information extraction, and more recently recognition of textual entailment. These tasks are not really easy, but the field was able to formulate them in such a way that a limited form of inference could be performed, measured/evaluated, and rewarded. Task formulations are an important method of guiding a research field: Summarization, another possible test-bed for text understanding, has gone an entirely different way in the past 15 years, where almost no advance towards reasoning has been made and where the most successful methods are instead based on statistical interactions between keywords alone.

IBM Watson is a recent successful example of how far we can go with NL-extracted knowledge and reasoning. The system relies on shallow inference, combined with many expert micro-models (time, prices, location, sports, music, culture). Its success in beating humans in the Jeopardy game show proofs that statistical and shallow approaches can be pushed very far, and that it can cover many types of knowledge. But can this approach be pushed infinitely far? In other words, is it just a matter of scale alone? What makes this question hard to answer is that we don't know exactly how much knowledge is needed (and for exactly which task?), versus how much we can obtain with foreseeable methods.

Discussing failures of IBM Watson, we talked about whether they were "only" due to the specific task (e.g., the limited reasoning time the Jeopardy setting) and some random knowledge gaps, or whether the lack of a more symbolic intermediate language was fundamentally hindering this kind of approach. For instance, Watson's time model is powerful enough to often influence decisions in the right way, but it was not able to overrule some very wrong inferences.

Another question concerns the fact that in a system without an intermediate representation, we have no access to interpret internal reasoning steps beyond a tracing mechanism in the form of linkages and statistical reasoning steps.

Designing such an intermediate language that could store extracted and some newly generalized knowledge is crucial and would be very hard. (Some might even say impossible in principle.) We speculated that it could be (light) symbolic, logic-based, or natural language-based. It should support explanation generation. It should be designed in such a way that it can be at least back-translated into some NL statements that are at least semi-understandable to humans.

Another point to consider is that an informal NL-based representation makes knowledge acquisition easier. This means it might also be possible to solicit knowledge from (naive) humans, e.g. by crowd-sourcing. Aiming at soliciting "folk-knowledge" rather than exact scientific knowledge, we would not need the extensive knowledge engineering from experts needed for the early expert systems in AI.

What about the kind of reasoning we would need in the next generation of reasoners? Quite likely, if it is to be useful in large-scale environments, it cannot be fully formal, and it cannot be rigid. Many well-researched types of relaxed inference have been developed in the field of AI and could be used (Bayesian, neural, case-based etc). Some truth conditional constraints can be and should be relaxed, but not all. While a system might consider probabilistically how likely it is that somebody who was insulted reacts in way X or Y, or that "a person's lifespan might be 120 years" (with low likelihood), it should absolutely blocking any inference relying on overestimations of a human lifespan by 400%. It would be desirable for the next generation of automatic reasoning to have this ability. As a side-point here, we also thought that human reasoning (folk-reasoning) could also establish gold standards for the kind of inference that might be useful for next-generation reasoning systems.

Even though our break-out group probably had a strong NLP bias due to group composition, there are also other possibilities of acquiring world knowledge which do not use written natural language at all. Non-language world knowledge could be captured by sensors, such as vision videos, which might allow for system to generalize that rain is wet or how people tend to behave in crowded shopping streets on a Saturday afternoon (from CCTVs). These are the kinds of facts about the world that an embedded intelligence, human or otherwise, learns by experience.

## 4.4 Argument and Argumentation Quality

*Henning Wachsmuth (Bauhaus-Universität Weimar, DE)*

**Joint work of** Wolf-Tilo Balke; Graeme Hirst; Yufang Hou; Vindokumar Prabharakan; Ruty Rinott; Christian Stab; Henning Wachsmuth

Debating technologies and related systems seek to use and provide single arguments and/or complete argumentations of high quality. As discussed in the talk "Analysis of Stance and Argumentation Quality", different quality dimensions have been investigated in research, whereas a common understanding of argument and argumentation quality is missing so far. Accordingly, the question of how to assess such quality has come up several times during the seminar. As a consequence, a working group has been formed during the seminar to coordinate future research on argument and argumentation quality.

Within the seminar, the concrete objective of the working group was to take a first step towards a taxonomy of quality dimensions. This resulted in the notion of a *contextualization* of quality assessment. The underlying made observation is that the context of assessing quality impacts what quality dimensions are seen as important. While several context factors have been discussed in the working group, the four that all agreed upon refer to the pursued *goals*, the used *medium*, the considered *granularity*, and the *view* on it. The working group identified the following values to exist for the four context factors:

- *Goals:* persuasion, deliberation
- *Medium:* text, speech, embodied
- *Granularity:* argument, argumentation
- *View:* monological, dialogical

In addition, the *mode* of persuasion was proposed as a fifth context factor:

- *Mode:* logos, pathos, ethos

However, a critical discussion of a case study revealed that the mode does not fully match the idea of context factors. The case study refers to a specific *contextualization*, i.e., the choice of one value for each context factor. In particular, the working group collected and clustered the following important quality dimensions for the contextualization "persuasive, textual, monological argument":

1. Coherence, adherence, clarity, conciseness, lexical quality
2. Logical correctness, completeness, consistency, validity of reasoning
3. Soundness
4. Truth, defendability, credibility, honesty
5. Convenience, comfort, easy accessibility
6. Relevance, utility, usefulness

As can be seen, the first cluster focuses on representational aspects, which distinguishes it from all other clusters. Cluster 2 clearly summarizes logos aspects, but cluster 4 and cluster 5 relate, at least partly, to ethos and pathos, respectively. Soundness forms its own cluster 3, as it captures characteristics of both cluster 2 and cluster 4. The quality dimensions of the remaining cluster 6, finally, may be somehow affected by all or most other dimensions. In contrast, cluster 6 addresses a different aspect of arguments, namely, whether the arguments help to achieve the purpose they are made for.

As implied above, the presented results are meant only as a first step towards a taxonomy of quality dimensions and towards a common understanding of argument and argumentation quality. Many of the findings sketched here are subject to further investigation. Also, their intersection and compliance with related work from argumentation theory still needs to be clarified. Besides, we explicitly point out that the context factors, their possible values, and the associated quality dimensions might neither be complete nor optimally defined so far. Similarly, the newly introduced terms (e.g., *context factor*) should be seen as preliminary only. That being said, the collaboration of the working group is still ongoing. Important results of this collaboration are planned to be published as soon as available.

## References

**1** J. Anthony Blair. *Relevance, Acceptability and Sufficiency Today*. In Groundwork in the Theory of Argumentation, pp. 87–100. 2012.

# 5 Panel discussions

## 5.1 Unshared Task Session

*Ivan Habernal (TU Darmstadt, DE), Iryna Gurevych (TU Darmstadt, DE), and Christian Stab (TU Darmstadt, DE)*

Shared tasks have been playing a major role in boosting research in many NLP fields. The availability of shared annotated data, clear evaluation criteria, and the presence of several competing systems allow for a fair direct comparison and overall progress tracking which in turn foster future research. However, argumentation mining, as an evolving research field, suffers not only from large data sets but also from a missing unified perspective on the tasks to be fulfilled as well as their evaluation. Given the variety of argumentation models, various argumentative genres and registers, granularities (i.e, micro-argumentation and macro-argumentation), dimension of argument (logos, pathos, ethos) and the overall social context of persuasion, the goal of defining a widely accepted task requires a substantial agreement, driven by empirical decision making. In this session, we thus conducted the so-called unshared task, whose goal is to come up with own definion of the task being tackled, the annotation scheme, self-assessment of its strenghts and weaknesses, and requirements for expertise of potential annotators, given only the plain text data. We experimented with five different registers (debate transcripts, forum posts, opinionated newswire articles, on-line discussions attached to an article, and pro-con debate portals) split among 28 participants. After an initial individual session, groups were established with respect to the data type and brainstormed their findings. Finally, a plenary discussion was held with presentation of the main findings from the groups, with emphasis on the criteria introduced above. Some of the results emerging from the discussion revealed that the claim-premise scheme is applicable to some data, but there is a need for capturing also the pragmatic layer (such as the activity, purpose, roles), user interactions, attribution, or patterns of debating. Outputs from this pilot experiment will partly serve for the upcoming unshared task at Argumentation Mining workshop at ACL 2016. A follow-up session was devoted to an active participation in playing Argotario – serious game for learning argument writing, component identification, and stance recognition [1]. The goal of this session was to benchmark the application under real-world conditions, with multiple players at the time. During 45 minutes of playing time, 12 users composed about 40 arguments, and answered about 60 questions in the two remaining game rounds. Several directions for future development were identified, such as player vs. player mode, or assessing argumentation quality.

### References

1 Raffael Hannemann. *Serious Games for large-scale Argumentation Mining*. Master Thesis, Technische Universität Darmstadt, https://www.ukp.tu-darmstadt.de/publications/details/?tx_bibtex_pi1[pub_id]=TUD-CS-2015-0108, 2015.

## 5.2 Debate and Argument Visualization

*Brian Plüss (The Open University – Milton Keynes, GB)*

In this discussion session, we consider issues around the visualisation of argumentation and debate. The topics proposed for discussion include the role of visualisations in argumentation research and communication; the possible targets of argument and debate visualisations (e.g., experts, the general public); what and how much should be visualised depending on target audiences; effective ways to evaluate visualizations; etc.

## Participants

- Khalid Al-Khatib
Bauhaus-Universität Weimar, DE
- Jens Allwood
University of Göteborg, SE
- Carlos Alzate
IBM Research – Dublin, IE
- Wolf-Tilo Balke
TU Braunschweig, DE
- Yonatan Bilu
IBM – Haifa, IL
- Elena Cabrio
INRIA Sophia Antipolis – Méditerranée, FR
- Claire Cardie
Cornell University, US
- Walter Daelemans
University of Antwerp, BE
- Ido Dagan
Bar-Ilan University – Ramat Gan, IL
- Anette Frank
Universität Heidelberg, DE
- Norbert Fuhr
Universität Duisburg-Essen, DE
- Iryna Gurevych
TU Darmstadt, DE
- Ivan Habernal
TU Darmstadt, DE
- Graeme Hirst
University of Toronto, CA
- Yufang Hou
IBM Research – Dublin, IE
- Eduard H. Hovy
Carnegie Mellon University – Pittsburgh, US
- Christoph Lofi
TU Braunschweig, DE
- Marie-Francine Moens
KU Leuven, BE
- Brian Plüss
The Open University – Milton Keynes, GB
- Vinodkumar Prabhakaran
Stanford University, US
- Chris Reed
University of Dundee, GB
- Nils Reiter
Universität Stuttgart, DE
- Ruty Rinott
IBM – Haifa, IL
- Hinrich Schütze
LMU München, DE
- Noam Slonim
IBM – Haifa, IL
- Christian Stab
TU Darmstadt, DE
- Manfred Stede
Universität Potsdam, DE
- Benno Stein
Bauhaus-Universität Weimar, DE
- Simone Teufel
University of Cambridge, GB
- Anita De Waard
Elsevier Labs – Jericho, US
- Henning Wachsmuth
Bauhaus-Universität Weimar, DE