

# Towards a Unified Model of Scholarly Argumentation

Khalid Al-Khatib<sup>\*1</sup>, Anita de Waard<sup>\*2</sup>, Dayne Freitag<sup>3</sup>,  
Iryna Gurevych<sup>\*4</sup>, Yufang Hou<sup>\*5</sup>, and Harrison Scells<sup>†6</sup>

- 1 University of Groningen, NL. [khalid.alkhatib@rug.nl](mailto:khalid.alkhatib@rug.nl)
- 2 Elsevier – Jericho, US. [a.dewaard@elsevier.com](mailto:a.dewaard@elsevier.com)
- 3 SRI International, US. [daynefreitag@sri.com](mailto:daynefreitag@sri.com)
- 4 TU Darmstadt, DE. [gurevych@cs.tu-darmstadt.de](mailto:gurevych@cs.tu-darmstadt.de)
- 5 IBM Research – Dublin, IE. [bnuxiaofang@gmail.com](mailto:bnuxiaofang@gmail.com)
- 6 Universität Leipzig, DE. [harry.scells@uni-leipzig.de](mailto:harry.scells@uni-leipzig.de)

---

## Abstract

This report summarizes the outcomes of the Dagstuhl Seminar 22432: “Towards a Unified Model of Scholarly Argumentation.” The purpose of this Seminar was to enable robust advances in argumentation technology by collecting and collaborating on use cases in scholarly and biomedical discourse and working on a foundational model for argumentation in science and healthcare. Most importantly, the seminar served to develop a multidisciplinary, international research community devoted to building and maintaining principles, tools, and models for studying scholarly argumentation. Over the course of the seminar week, the seminar laid the foundation of a shared formalism, illuminated important scholarly use cases for argumentation modeling, and identified directions for future exploration.

**Seminar** October 23–28, 2022 – <http://www.dagstuhl.de/22432>

**2012 ACM Subject Classification** Computing methodologies → Artificial intelligence; Theory of computation; Computing methodologies → Machine learning

**Keywords and phrases** Argument mining, Argument modeling, Scholarly discourse

**Digital Object Identifier** 10.4230/DagRep.12.10.175

## 1 Executive Summary

*Khalid Al-Khatib (University of Groningen, NL, [khalid.alkhatib@rug.nl](mailto:khalid.alkhatib@rug.nl))*

*Anita de Waard (Elsevier-Jericho, US, [a.dewaard@elsevier.com](mailto:a.dewaard@elsevier.com))*

*Iryna Gurevych (TU Darmstadt, DE, [iryana.gurevych@tu-darmstadt.de](mailto:iryana.gurevych@tu-darmstadt.de))*

*Yufang Hou (IBM Research-Dublin, IE, [yhou@ie.ibm.com](mailto:yhou@ie.ibm.com))*

**License**  Creative Commons BY 4.0 International license

© Khalid Al-Khatib, Anita de Waard, Iryna Gurevych, and Yufang Hou

## Background

Argumentation is prevalent in scientific discourse and critical to scientific progress. Recent efforts have attempted to identify and model argumentative structures in scholarly discourse from different perspectives. Within the domain of scientific literature analysis, computational approaches to argumentation have followed the route of discourse modeling by identifying relations between spans and clauses encoding rhetorical structures (e.g., premises and conclusions), or as typed turns in community debate (e.g., supports or attacks). Another thread of research, often applied to biomedical literature, focuses on capturing functional

---

\* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Towards a Unified Model of Scholarly Argumentation, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 175–206

Editors: Khalid Al-Khatib, Anita de Waard, Dayne Freitag, Iryna Gurevych, Yufang Hou, and Harrison Scells



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

discourse at different levels of granularity, such as objectives, methods, results or scientific claims, and their relation to reported evidence. Most work adopts a corpus perspective, either highlighting the role of sentences or phrases within the scientific discourse or aligning claims across documents, and using citations to construct claim-evidence networks that summarize the state of knowledge in a field. Within the health sciences, argumentative structures have been used to automate the production of systematic reviews by identifying key actionable knowledge elements from collections of clinical reviews, case studies, and research papers. For an overview of previous work, see e.g. [1].

Despite these varied efforts and the clear practical importance of the work, there is lack of consensus on how scientific argumentation should be formalized. For instance, it remains unclear whether formalisms popular in non-scientific domains apply to scientific discourse, and whether a single formalism can adequately support argumentation research in diverse disciplines such as biology, chemistry, materials science, medical research and computer science. This lack of consensus manifests in a dearth of shared reference corpora, which are needed to advance research into computational treatments of scientific argumentation. It has also led to the absence of an operational theory for defining argumentative components in scholarly text.

## Goals

Our Dagstuhl Seminar, titled *Towards a Unified Model for Scholarly Argumentation*, sought to further the emergence of this missing consensus. Specifically, the seminar objectives included:

- Enabling robust advances in argument technology by collecting and working on use cases in scholarly and medical discourse;
- Starting the development of a foundational model for argumentation in science and healthcare;
- Laying the groundwork for a multidisciplinary community devoted to building and maintaining principles, tools, and models to identify key components in scholarly argumentation.

## Outcomes

The seminar was attended by scientists at different levels of seniority and from a variety of research backgrounds. Some participants have made the computational modeling of argumentation or the scholarly literature the central focus of their careers. Others were drawn to the seminar through their work on applications in adjacent problem areas. Ultimately, all emerged with a sense that important bonds of shared interest had formed, fostered by several seminar outcomes.

## Knowledge Baseline

A shared understanding of the problem space was obtained, through a series of keynotes and panel discussions on theory, models, tools, and available corpora. These are described in greater detail in this report, in Section 3. In particular, two introductory talks summarized the state of the art in argument modeling (3.1) and computational argument mining (3.2).

Five further plenary talks described different use cases where argument identification can support NLP tasks:

- using scientific discourse to understand and measure the impact of scholarly contributions (3.3);
- using argument modeling to generate discourse (3.4);
- generating scholarly documents using argument structures (3.5);
- interpreting a fortiori arguments (3.6);
- synthesizing evidence from text to support public policy (3.7).

A series of eleven flash talks covered a host of other efforts, presenting corpora, tools, and relevant applications, such as document understanding, extracting high-level claims, and identifying fallacious and persuasive elements in scholarly texts (Section 4).

### Problem Elucidation

At the beginning of the workshop, the group identified several important focus areas that then became the subject of breakout group deliberation over the course of the week. All materials, including the full program, slides, summaries of the breakout sessions and code and corpora submitted can be found on the workshop Google Drive at <https://bit.ly/TUMSA22>.

- *Foundations* (Section 5.1). A subgroup of participants discussed a shared argumentation model, based on the various proposals presented during the plenary sessions. The group debated and wrote a first-order consensus of these varying views, which can be used for further development of a foundational model of scholarly argumentation.
- *Domains* (Section 5.2). This working group pursued a comparison of argumentation in different scholarly domains. A methodology was delineated for how to annotate argumentation across domains while reducing the need for domain experts.
- *Argument Quality* (Section 5.3). This working group explored how argumentation quality can be evaluated, and defined a series of questions to assess this. Additionally, the group members contributed an open-source tool to perform the evaluation of argumentation quality, which can be further developed to support this task.
- *Community Dialogue* (Section 5.4). This working group looked at how argument structure can support an important editorial task, namely to decide on an accept or reject decision for a submitted manuscript, based on a number of peer reviews. The group developed a corpus of dialogues that simulate how a meta-reviewer asks questions about a document that has received a number of reviews, which can be used in future work in this domain.

### Community Formation

Building on the connections developed during the seminar, a series of collaborations have been fostered, and thoughts on how to proceed with this work through a multidisciplinary lens have been put forth. Multiple new collaborations have been formed as an outcome of this week, in some cases centered on new tools and research corpora first conceived in the workshop.

### Next Steps

This Dagstuhl Seminar brought together a multi-disciplinary, international, and diverse community of researchers from academia and industry to discuss scholarly argumentation. Much argumentation occurred, during and after presentations, in breakout groups, during

the social events spread out through the week, and long into the night. Necessarily, this is only the beginning of a conversation that will unfold over the coming years, one that will ultimately produce a shared model of scholarly argumentation and a set of concrete research tasks and important new use cases.

We hope that this seminar was the first in a series of events devoted to this topic, that this inaugural event proves pivotal in the formation of a cohesive research community addressing a problem with large practical ramifications. This report can hopefully contribute to accelerate work in this area, by offering a summary of current efforts, and a number of interesting problems to work on.

### References

- 1 Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. 2021. Argument Mining for Scholarly Document Processing: Taking Stock and Looking Ahead. In Proceedings of the Second Workshop on Scholarly Document Processing, pages 56–65, Online. Association for Computational Linguistics.

## 2 Table of Contents

### Executive Summary

*Khalid Al-Khatib, Anita de Waard, Iryna Gurevych, and Yufang Hou* . . . . . 175

### Introductory Talks

An introduction to Models of Argumentation  
*Graeme Hirst and Chris Reed* . . . . . 181

Computational Argumentation in Scholarly Discourse  
*Khalid Al-Khatib and Henning Wachsmuth* . . . . . 181

Towards Automatically Understanding and Measuring the Contributions of Scientific Work  
*Maria Liakata* . . . . . 181

The Role of Text Generation in Argumentation  
*Smaranda Muresan* . . . . . 182

InterText: Modeling Text as a Living Object in Cross-Document Context  
*Iryna Gurevych* . . . . . 182

Formalizing and Generating the Structure of Scholarly Papers  
*Eduard Hovy* . . . . . 183

Towards Automatic Interpretation of A Fortiori Arguments  
*Simone Teufel* . . . . . 183

Using the Claim Framework to Inform Public Policy  
*Ryan Wang* . . . . . 184

### Flash Talks

Narrative Structures in Scientific Documents  
*Wolf-Tilo Balke* . . . . . 184

Argumentation in Biochemistry Articles  
*Robert Mercer* . . . . . 185

Linking Computational Argumentation to Information Quality  
*Davide Ceolin* . . . . . 186

Building Computational Models to Understand Scholarly Documents  
*Yufang Hou* . . . . . 186

PEER – Collaborative Lightweight Argument Annotation  
*Nils Dycke* . . . . . 187

Towards Constructive Conversations  
*Andreas Vlachos* . . . . . 187

Expressing High-Level Scientific Claims with Formal Semantics  
*Davide Ceolin* . . . . . 188

Argumentation, Persuasion, Propaganda, and More  
*Preslav Nakov* . . . . . 188

Fallacies in Political Argumentation  
*Serena Villata* . . . . . 189

Communicating Scientific Work with the Public through Dialogue Initiative <i>Milad Alshomary and Smaranda Muresan</i> . . . . .	189
BAM: Benchmarking Argument Mining on Scientific Documents <i>Florian Ruosch</i> . . . . .	190
<b>Working Groups</b>	
Foundations of Scholarly Argumentation <i>Elena Cabrio, Graeme Hirst, Eduard Hovy, Maria Liakata, Robert Mercer, Smaranda Muresan, Preslav Nakov, Chris Reed, Florian Ruosch, Simone Teufel, Serena Villata</i>	190
Cross-domain Argumentation Model for Scholarly Argumentation <i>Khalid Al-Khatib, Fengyu Cai, Dayne Freitag, Daniel Garijo, Benno Stein, Henning Wachsmuth</i> . . . . .	196
Evaluation of Argument Quality <i>Yufang Hou, Tobias Mayer, Domenic Rosati, Harrisen Scells, Ferdinand Schlatt, Simone Teufel, Ryan Wang</i> . . . . .	200
Scholarly Argumentation as a Community Dialogue <i>Wolf-Tilo Balke, Andreas Vlachos, Davide Ceolin, Milad Alshomary, Nils Dycke, Sukannya Purkayastha, Iryna Gurevych, Anne Lauscher, Tilman Beck</i> . . . . .	202
<b>Participants</b> . . . . .	206

## 3 Introductory Talks

### 3.1 An introduction to Models of Argumentation

*Graeme Hirst (University of Toronto, CA) and Chris Reed (University of Dundee, UK)*

License © Creative Commons BY 4.0 International license  
© Graeme Hirst and Chris Reed

We reviewed the fundamental concepts of arguments and argumentation, including the basic elements of arguments, the types of argument structures, and the types of attacks on arguments. We introduced the idea of argumentation as a dialogue game, and the conditions required of a well-formed argument. We outlined the Toulmin model of argumentation, and explain the concept of argumentation schemes as templates for arguments.

### 3.2 Computational Argumentation in Scholarly Discourse

*Khalid al-Khatib (University of Groningen, NL) and Henning Wachsmuth (Leibniz Universität Hannover, DE)*

License © Creative Commons BY 4.0 International license  
© Khalid Al-Khatib and Henning Wachsmuth

Computational argumentation deals with the computational analysis and synthesis of natural language arguments. In this tutorial talk, we provided an overview of computational argumentation from a natural language processing (NLP) perspective, and we reviewed the state of the art of computational argumentation in scholarly discourse. Starting from the basics of human argumentation, the first part of the talk introduced the central tasks of argument mining, argument assessment, and argument generation. We then looked at the latest trends for these tasks considering audience-specific argument quality assessment and knowledge encoding during argument generation. In the second part, we concentrated on scholarly discourse organizing existing research based on the domains being tackled and the argument models built on. Most existing work addresses the creation of new corpora for scholarly documents and the mining of their argumentative structure. We discussed the main envisioned applications of computational argumentation in scholarly discourse and the challenges towards these.

### 3.3 Towards Automatically Understanding and Measuring the Contributions of Scientific Work

*Maria Liakata (The Alan Turing Institute – London, UK & Queen Mary University of London, UK)*

License © Creative Commons BY 4.0 International license  
© Maria Liakata

Researchers have been working on the automatic extraction of information from scientific articles for over two decades. A key aspect in this line of research is capturing how scientists discuss their work, the scientific discourse. In my talk I gave a brief overview of early work on identifying the scientific discourse and how this can improve downstream tasks involving

the extraction of information from the scientific literature. I then showed a number of neural approaches to capturing scientific argumentation in a multi-task learning setting. I also presented recent work on the relation between the scientific discourse and the way it is represented in the news through cross-document cross-domain coreference between scientific articles and news and press releases that refer to the scientific articles, as a step towards understanding the more comprehensive (non-academic) impact of scientific work.

### 3.4 The Role of Text Generation in Argumentation

*Smaranda Muresan (Columbia University – New York City, USA)*

License  Creative Commons BY 4.0 International license  
© Smaranda Muresan

Large-scale language models based on transformer architectures, such as GPT-3 or BERT, have advanced the state of the art in Natural Language Understanding and Generation. However, even though these models have shown impressive performance for a variety of tasks, they often struggle with reasoning and modeling implicit meaning, which are required for understanding and generating argumentative text. In this talk, I presented some of our recent work on text generation models for argumentation. There are several challenges we have to address to make progress in this space: 1) the need to model commonsense knowledge; 2) the lack of large training datasets. I discussed our proposed theoretically-grounded knowledge-enhanced text generation models for enthymeme reconstruction and for recognizing argument fallacies. I concluded by discussing opportunities and remaining challenges for neural text generation systems for argumentation.

### 3.5 InterText: Modeling Text as a Living Object in Cross-Document Context

*Iryna Gurevych (Technical University Darmstadt, DE)*

License  Creative Commons BY 4.0 International license  
© Iryna Gurevych

The ability to find and interpret cross-document relations is crucial in many fields of human activity, from social media to collaborative writing. While natural language processing has made tremendous progress in extracting information from single texts, a general NLP framework for modeling interconnected texts including their versions and related documents is missing. The talk reported on our ongoing efforts to establish such a framework. We addressed several challenges related to this. First, NLP has an acute need for diverse data to model cross-document tasks. We discussed our new, ethically sound data acquisition strategies and present unique cross-document datasets in the scientific domain, along with a generic data model that can capture text structure and cross-document relations in heterogeneous documents. Second, we reported on a study that instantiates our framework in the domain of scientific peer reviews. Finally, we highlighted our vision for cross-document computational argument analysis instantiating the InterText framework for analyzing arguments across documents. Our results pave the way to move NLP forward towards more human-like interpretation of text in the context of other texts.



### 3.6 Formalizing and Generating the Structure of Scholarly Papers

*Eduard Hovy (Carnegie Mellon University – Pittsburgh, USA & University of Melbourne, AU)*

License © Creative Commons BY 4.0 International license  
© Eduard Hovy

As robust single-sentence generation in response to a prompt is more or less a solved issue now, and the controlled production of a coherent longer text is very much under investigation, one can wonder: what would it take to automatically generate a scholarly paper? In this talk I described (1) the representation in a structured form of the scholarly content; (2) the genre-oriented information required in scholarly discourse; (3) how to compose the first kind of information with the second using a typical modern neural network approach to argumentation structure. Topic (1) describes frameworks that can serve as templates for scholarly information; topic (2) outlines some rhetorical functions that information must be cast into to produce the appropriately structured scholarly genre; and topic (3) surveys various approaches and architectures to perform the requisite text planning.

### 3.7 Towards Automatic Interpretation of A Fortiori Arguments

*Simone Teufel (University of Cambridge, UK)*

License © Creative Commons BY 4.0 International license  
© Simone Teufel

In this talk, I reported on work by my PhD student Olesya Razuvaevskaya. Her starting point was the restoration of premises in mini-arguments, whereby we wanted the generated premise to be guaranteed to be logically valid, as well as objectively explainable. We concentrated on the phenomenon of A fortiori logic, a logically valid reasoning pattern that has been known since ancient times and that is very frequent in day-to-day language use. Starting from sentences containing the phrase “let alone”, our analysis uses the fact that two situations are described and compared in terms of their likelihood. This simple fixed structure allows us to isolate the underlying logic to a single principle per argument, with just a few parameters necessary for explaining each case. The cases we consider are a) two quantities are concerned; b) the difference in likelihood concerns specificity; c) one of the situations described is a precondition of the other, and d) some underlying resource not mentioned in the text is required to explain the difference in likelihood. The d) cases require deeper reasoning. I also described key points of Olesya’s implementation of a system for the automatic partial interpretation of a system for a fortiori interpretation. The implementation uses standard neural sequence analysers and masked and unmasked transformers to provide a modular, pipelined analysis of three core aspects of the analysis.

### 3.8 Using the Claim Framework to Inform Public Policy

Ryan Wang (*University of Illinois Urbana-Champaign – Urbana, USA*)

License  Creative Commons BY 4.0 International license  
© Ryan Wang

In this talk, I discussed the Claim Framework [1] and its application in evidence-based policymaking. The Claim Framework is concerned with the identification and representation of five kinds of scientific claims: the explicit claim, the implicit claim, observation, correlation, and comparison. An explicit claim consists of two entities connected by a relationship term that indicates a change observed in the experiment. An implicit claim similarly has two entities but the relationship between those is expressed in a more implicit manner. Unlike explicit and implicit claims, an observation identifies a change and the entity impacted by the change while leaving out the entity that causes the change. A claim that describes a correlation between two entities is a correlation. Finally, a comparison is a comparative construction where two entities are compared on a common ground. Taken together, the Claim Framework offers a principled means of extracting and organizing scientific claims that can be of great value to policymakers. [2] provides an example that uses the Claim Framework to automatically extract supporting, neutral, and refuting evidence of cell death and proliferation from biomedical abstracts with the aim of accelerating the otherwise time-consuming process of chemical risk assessment.

#### References

- 1 Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *J. Biomed. Inform.* 43, 2 (April 2010), 173–189. <https://doi.org/10.1016/j.jbi.2009.11.001>
- 2 Catherine Blake and Jodi A. Flaws. 2021. Using semantics to scale up evidence-based chemical risk-assessments. *PLoS ONE* 16, 12, Article e0260712 (15 Dec. 2021), 24 pages. <https://doi.org/10.1371/journal.pone.0260712>

## 4 Flash Talks

### 4.1 Narrative Structures in Scientific Documents

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

License  Creative Commons BY 4.0 International license  
© Wolf-Tilo Balke

From early on, narratives have been used as an essential means to convey information and knowledge in a form that is close to human communication and sense making. Moreover, references to archetypical narratives, such as David vs. Goliath, can also transport a set of connotations beyond the actual story allowing for a framing of information in the sense of speech acts. Facing today's flood of data and scientific results, data-driven narratives are thus an ideal way to make complex topics comprehensible, to make sense of certain events, or to assess the plausibility of given narratives or lines of arguments. However, these features are rarely used in information systems today. In particular, most of the current work on narratives is limited to representing structural properties such as story or plot graphs/plot units, event chains, or representations of entities and events without exploiting the deeper meaning of narratives. We explore narratives in the sense of logical overlays over

heterogeneous knowledge repositories, such as knowledge graphs, linked open data sources, document collections, or even concrete datasets. In its simplest form, a narrative then is a directed graph consisting of entities, events, and literals as nodes. Narrative edges describe the flow of the modeled events, i.e. on the one hand the semantic interaction between events and entities and on the other hand the respective types of interaction by suitable edge labels (e.g., in the causal or temporal sense). Essential for the expressive power of this overlay model is that edges of a narrative must always be bound against underlying knowledge repositories. In particular, this allows the plausibility of each edge to be evaluated against a given set of trusted repositories. Of course, this also means that the information in the underlying repositories needs to be carefully extracted with respect to classical dimensions of data quality, such as correctness, completeness, or validity.

## 4.2 Argumentation in Biochemistry Articles

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

**License** © Creative Commons BY 4.0 International license  
© Robert Mercer

**Joint work of** Eli Moser, Robert Mercer

**Main reference** Eli Moser, Robert E. Mercer: “Use of Claim Graphing and Argumentation Schemes in Biomedical Literature: A Manual Approach to Analysis”, in Proc. of the 7th Workshop on Argument Mining, pp. 88–99, Association for Computational Linguistics, 2020.

**URL** <https://aclanthology.org/2020.argmining-1.10>

This talk presented our contributions to argumentation in the experimental life sciences, scholarly biochemistry articles, in particular. Biomedical articles found in PubMed divide naturally into two classes: clinical and experimental. In the experimental class two types of articles have been or are being studied: genetics and biochemistry. With evidence from five biochemistry articles, the argumentation schemes that Green [2] has proposed for genetics articles transfer to biochemistry. We have studied the argumentation graphs that can be produced from the premises and claims in these articles and suggest an argumentation scheme hierarchy that is found therein. Biochemistry articles are structured in the IMRaD style (Introduction, Methods, Results, and Discussion). In work that is complementary to the well-known Argumentation Zoning model [4], Kanoksilapathum [3] has proposed rhetorical moves for each of these four sections. Providing computational models to identify these moves is ongoing work. In addition to the argumentation structure that exists in the main body of an article, titles with finite verbs strongly indicate the main claim of the article [1]. And structured abstracts provide similarly organized summaries of each of the four IMRaD sections. Work proceeds to connect Rhetorical Structure Theory to the argumentation schemes found in scholarly biochemistry articles with the ultimate goal of automating the identification of the schemes.

### References

- 1 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.
- 2 Nancy Green. 2015. Identifying argumentation schemes in genetics research articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21.
- 3 Budsaba Kanoksilapatham. 2005. Rhetorical structure of biochemistry research articles. *English for Specific Purposes*, 24(3):269–292.
- 4 Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

### 4.3 Linking Computational Argumentation to Information Quality

*Davide Ceolin (Centrum Wiskunde & Informatica, Amsterdam, NL)*

License  Creative Commons BY 4.0 International license  
© Davide Ceolin

The logical and argument structure of information items can be an indicator of their information quality. In this talk, we presented a transparent pipeline to automatically mine and reason on arguments from information items [1, 2]. We evaluate how such argument-based analyses reflect on information quality by comparing argument-based assessments with quality assessments considering diverse aspects of quality (e.g., veracity, precision, completeness). The pipeline we propose combines diverse components based on machine learning, symbolic reasoning, and human computation. We evaluate the impact of diverse implementations of these components and test the pipeline on a dataset of product reviews. We plan to extend this pipeline to analyze scholarly documents in the future.

#### References

- 1 Davide Ceolin, Giuseppe Primiero, Jan Wielemaker, and Michael Soprano. 2021. Assessing the Quality of Online Reviews Using Formal Argumentation Theory. In *Web Engineering: 21st International Conference, ICWE 2021, Biarritz, France, May 18–21, 2021, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 71–87. [https://doi.org/10.1007/978-3-030-74296-6\\_6](https://doi.org/10.1007/978-3-030-74296-6_6)
- 2 Ceolin, D, Primiero, G, Soprano, M, & Wielemaker, J. (2022). Transparent assessment of information quality of online reviews using formal argumentation theory. *Information Systems*, 110, 102107.1–102107.14. doi:10.1016/j.is.2022.102107

### 4.4 Building Computational Models to Understand Scholarly Documents

*Yufang Hou (IBM Research Europe – Dublin, IE)*

License  Creative Commons BY 4.0 International license  
© Yufang Hou

The accumulated scientific knowledge is the foundation upon which informed decision making is built, with huge impact across a wide range of critical applications. In this talk, I gave a short overview of my recent work on information extraction and natural language generation on scholarly documents, including interactive document2slides generation [1], scientific leaderboards construction [2], NLP TDM knowledge graph construction [3, 4]. Finally, I talked about our recent work on diachronic analysis of the NLP research areas, in which we developed a model to analyse NLP research areas and answer the following questions: (1) What is the general trend of a research area? (2) How is a research area influenced by other research concepts? (3) How do researchers argue about a specific research concept?

#### References

- 1 Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang and Nancy X.R. Wang. D2S: Automated Slide Generation With Query-based Text Summarization From Documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021)*, Online, 6–11 June 2021

- 2 Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Debasis Ganguly. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, 27 July-2 August 2019
- 3 Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Debasis Ganguly. TDMSci: A Specialized Corpus for Scientific Literature Entity Tagging of Tasks Datasets and Metrics. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021), Online, 19-23 April 2021
- 4 Ishani Mondal, Yufang Hou, Charles Jochim. End-to-End Construction of NLP Knowledge Graph. In Proceedings of Findings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021 Findings), Online, 1-6 August 2021 Association for Computational Linguistics.

## 4.5 PEER – Collaborative Lightweight Argument Annotation

*Nils Dycke (Technical University Darmstadt, DE)*

License  Creative Commons BY 4.0 International license  
© Nils Dycke

In this talk we introduced PEER, a collaborative, light-weight annotation tool for scholarly documents. The wide range of commercial tools for highlighting and commenting in PDFs (e.g. Google Docs, hypothesis, ...) cannot be used for scientific annotation studies at scale: they require uploading of confidential and potentially sensitive research data to public servers, and offer no mechanisms to manage, export or import annotation data. On the other hand, classical data annotation tools from the NLP community (e.g. Inception) require significant effort to set up for scholarly documents and, while being very feature-rich, they can be overwhelming to non-experts. To close this gap, we propose the PDF-annotation tool PEER, which unites the ease-of-use of highlighting and commenting software with the ease-of-access to NLP researchers of classical annotation tools. PEER offers a test bed for rapid prototyping different span annotation schemata and a lean study management interface. Users engage in their habitual process of highlighting and commenting in the annotation interface without the need for extensive familiarization with the tool. PEER comes as a ready-to-use web application and can be set up on local servers quickly. Hereby, we contribute towards the creation of new annotated datasets in the scholarly argumentation research.

## 4.6 Towards Constructive Conversations

*Andreas Vlachos (University of Cambridge, UK)*

License  Creative Commons BY 4.0 International license  
© Andreas Vlachos

**Joint work of** Christine De Kock, Youmna Farag, Georgi Karadzhov, Tom Stafford, Andreas Vlachos

In this talk I presented our work motivated by the question “What makes conversations among humans more constructive and how can we intervene to make them happen”. First, I discussed group decision-making in the context of the Wason Card Selection task [1], where we find that groups perform better than individuals, and, more interestingly, can reach a correct decision even if no one had it in the beginning of the conversation [2]. Following

this, I presented the Wikipedia disputes dataset [3] which has allowed us to examine how disagreements are resolved in the context of Wikipedia, the most successful large-scale collaborative project. Finally, I described our work on developing and evaluating a dialogue agent for exposing people to the opposing side of an argument [4].

#### References

- 1 Peter C Wason. 1968. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273281.
- 2 Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2022. DeliData: A dataset for deliberation in multi-party problem solving. <https://arxiv.org/abs/2108.05271>
- 3 Christine De Kock and Andreas Vlachos. 2021. I Beg to Differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5601–5613, Dublin, Ireland. Association for Computational Linguistics.
- 4 Farag, Youmna; Brand, Charlotte; Amidei, Jacopo; Piwek, Paul; Stafford, Tom; Stoyanchev, Svetlana and Vlachos, Andreas (2022). Opening up Minds with Argumentative Dialogues. In: *Findings of EMNLP (Empirical Methods in Natural Language Processing)*

### 4.7 Expressing High-Level Scientific Claims with Formal Semantics

*Davide Ceolin (Centrum Wiskunde & Informatica, Amsterdam, NL)*

License  Creative Commons BY 4.0 International license  
© Davide Ceolin

In this talk, we presented a method to express the content of high-level scientific claims using formal semantics in a systematic way [1]. Leveraging existing semantic formalisms, we developed the concept of “superpattern”, i.e., a formal representation of scientific claims corresponding to a conditional probability over logical formulas. Through this formalism, we can enable a full machine-understandable representation of scientific claims. The effectiveness of superpatterns has been evaluated both by effectively representing multiple claims from diverse scientific outlets, and by performing a user study that shows a high level of agreement among experts employing this technique.

#### References

- 1 Bucur, C-I, Kuhn, T, Ceolin, D & Ossenbruggen, JV 2021, “Expressing High-Level Scientific Claims with Formal Semantics”, arXiv, pp. 233-240. <https://doi.org/10.1145/3460210.3493561>

### 4.8 Argumentation, Persuasion, Propaganda, and More

*Preslav Nakov (Mohamed bin Zayed University of Artificial Intelligence – Abu Dhabi, AE)*

License  Creative Commons BY 4.0 International license  
© Preslav Nakov

We described the connection between argumentation, persuasion, and propaganda: what their goals are and what techniques they use. We presented a specific inventory of propaganda techniques and we show that they do appear in scholarly articles. We further discussed framing as well as the role of figures and citances in scholarly articles, esp. in the life sciences. Finally, we discussed ways to use text summarization techniques with the aim of producing a layman’s summary of a scholarly article.

## 4.9 Fallacies in Political Argumentation

*Serena Villata (Université Côte d'Azur, CNRS, Inria, I3S, France)*

**License** © Creative Commons BY 4.0 International license  
© Serena Villata

**Joint work of** Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, Serena Villata  
**Main reference** Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, Serena Villata:  
“Fallacious Argument Classification in Political Debates”, in Proc. of the Thirty-First International  
Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022,  
pp. 4143–4149, ijcai.org, 2022.  
**URL** <https://doi.org/10.24963/ijcai.2022/575>

First, I presented a novel annotated resource of 31 political debates from the U.S. Presidential Campaigns, where we annotated six main categories of fallacious arguments (i.e., ad hominem, appeal to authority, appeal to emotion, false cause, slogan, slippery slope) leading to 1628 annotated fallacious arguments. Second, I introduced this novel task of fallacious argument classification and I presented the neural architecture based on transformers we proposed. Our results show the important role played by argument components and relations in this task.

## 4.10 Communicating Scientific Work with the Public through Dialogue Initiative

*Milad Alshomary (Leibniz Universität Hannover, DE) and Smaranda Muresan (Columbia University – New York City, USA)*

**License** © Creative Commons BY 4.0 International license  
© Milad Alshomary and Smaranda Muresan

The gap between the scientific community and the public is growing. AI hype and distrust in science have become challenging issues nowadays. In our work, we aim to bridge this gap by first encouraging authors of scientific works to communicate their work to the public (e.g., journalists, non-experts, etc.). Instead of producing lay summaries, we hypothesize that the best form of communication is through dialogues, giving a space for both the authors and the public to construct an explanation and understanding of the subject matter jointly. Second, by studying the dialogical communication between these two parties, we can potentially provide assistant tools that can help authors sharpen their communication skills and (semi) automate the process of explaining scientific work to the public.

## 4.11 BAM: Benchmarking Argument Mining on Scientific Documents

*Florian Ruosch (Universität Zürich, CH)*

**License** © Creative Commons BY 4.0 International license  
© Florian Ruosch

**Joint work of** Florian Ruosch, Cristina Sarasua, Abraham Bernstein

**Main reference** Florian Ruosch, Cristina Sarasua, Abraham Bernstein: “BAM: Benchmarking Argument Mining on Scientific Documents”, in Proc. of the Workshop on Scientific Document Understanding co-located with 36th AAAI Conference on Artificial Intelligence, SDU@AAAI 2022, Virtual Event, March 1, 2022, CEUR Workshop Proceedings, Vol. 3164, CEUR-WS.org, 2022.

**URL** <http://ceur-ws.org/Vol-3164/paper5.pdf>

I presented BAM, a unified Benchmark for Argument Mining (AM): a method to homogenize both the evaluation process and the data to provide a common view in order to ultimately produce comparable results. Built as a four stage and end-to-end pipeline, the benchmark allows for the direct inclusion of additional argument miners to be evaluated. First, the system pre-processes a ground truth set used both for training and testing. Then, the benchmark calculates a total of four measures to assess different aspects of the mining process. To showcase an initial implementation of our approach, the procedure is applied and evaluates a set of systems on a corpus of scientific publications. With the obtained comparable results, we can homogeneously assess the current state of AM in this domain.

## 5 Working Groups

### 5.1 Foundations of Scholarly Argumentation

*Elena Cabrio (Université Côte d’Azur – Sophia Antipolis, FR)*

*Graeme Hirst (University of Toronto, CA)*

*Eduard Hovy (Carnegie Mellon University – Pittsburgh & University of Melbourne, AU)*

*Maria Liakata (The Alan Turing Institute – London, UK & Queen Mary University of London, UK)*

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

*Smaranda Muresan (Columbia University – New York City, USA)*

*Preslav Nakov (Mohamed bin Zayed University of Artificial Intelligence – Abu Dhabi, AE)*

*Chris Reed (leader) (University of Dundee, UK)*

*Florian Ruosch (Universität Zürich, CH)*

*Simone Teufel (University of Cambridge, UK)*

*Serena Villata (Université Côte d’Azur – Sophia Antipolis, FR)*

**License** © Creative Commons BY 4.0 International license

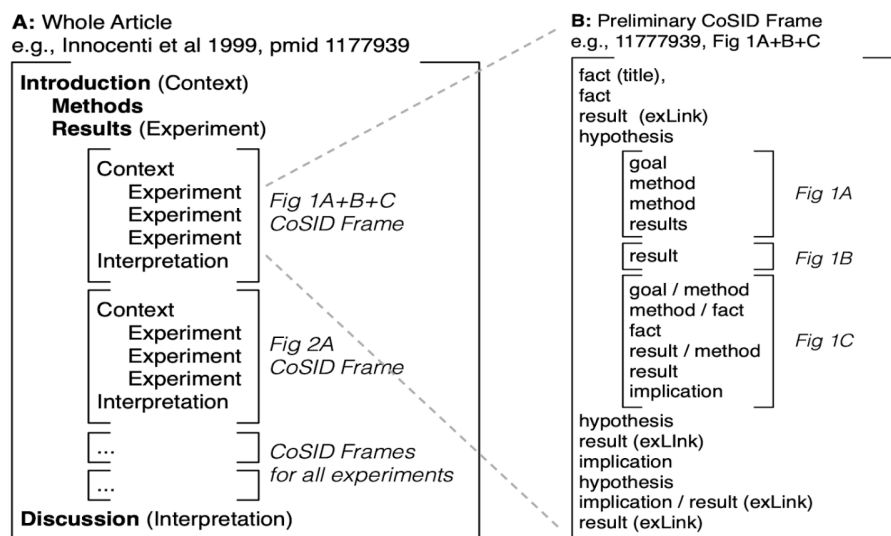
© Elena Cabrio, Graeme Hirst, Eduard Hovy, Maria Liakata, Robert Mercer, Smaranda Muresan, Preslav Nakov, Chris Reed, Florian Ruosch, Simone Teufel, Serena Villata

#### 5.1.1 Introduction

We develop a framework to represent scholarly argumentation presented in research papers. We attempt to make the framework compatible with as much existing work as feasible and strive not to introduce novelties that still need to be defined, verified, and generally accepted.

We take the approach that there exist different “genres” of scholarly papers, such as *Experiment Report*, *Mathematical Proof*, and *Research Survey*, among others. Each paper genre has a characteristic stereotypical structure. For example, an *Experiment Report*





■ **Figure 1** Frames describing a set of Experiments from [3].

paper includes the description of an experiment containing a hypothesis, methods employed, measurement procedure, and measured results, while a *Mathematical Proof* paper includes a claim and its proof. Different disciplines tend to prefer different genres.

Regardless of genre, a scholarly paper is an artifact containing text, images, and possibly data or software in which the author makes an argument in support of one or more claims. In addition to the core claim, an argument includes text to support the claim and text to refute contradictory claims. The internal structure of arguments consists of text blocks of various types that recursively contain smaller blocks, ending with (approximately) a clause as the basic unit. Typically each block fulfills a discourse function, such as *introduce* or *prove*. Blocks are related to units or other blocks in various ways, for example through coreference among units.

In this section of the report, we describe the most common blocks and their composition into typical scholarly paper structures. The bulk of the section provides sets of labels that characterize the types of blocks and the types of relations that hold between them.

### 5.1.2 Frames

We represent the internal (sub)structure of argument blocks using frames. A frame is a list of smaller blocks, each supporting a specific discourse function within the larger block. The label names the block's function. For example, the *Experiment Report* is the frame (the sequence of blocks) Hypothesis + Experiment + Conclusion, where Experiment consists of the frame Method + Measurement and Conclusion consists of Interpretation + Claim. One elaboration of an *Experiment Report* frame for Biomedicine was developed in [3], see Figure 1.

### 5.1.3 Annotation Layers

To define frames, we have to define their building elements: the labels. Each textual unit in a frame carries one or several labels. The “smallest” textual unit, the simple proposition, is approximately a clause.

At the outset, we note that some units in papers refer solely to domain objects and actions, which exist in the world independently of the author’s beliefs or argumentation, while others, including claims, hypotheses, proofs, etc., include the author’s beliefs and are used by the author to build the argument. We call the former the *Domain World* and the latter the *Rhetorical World*. In general, propositions from the latter world reflect (explicitly or implicitly) some aspect of the author’s opinion (beliefs about factuality or attitudes about desirability), while propositions from the former have no such connotation. Of course, additional worlds of annotation exist, notably the Evaluation World to capture readers’ assessments of the argument in the paper. This Evaluation World is the focus of the Evaluation Group (see Section 5.3).

Typically, a clause has one or more labels from each world, plus perhaps linkages to units elsewhere in the paper. Corresponding to these worlds, we annotate a paper at two separate layers, each world providing its own set of labels and assigning additional information to a textual unit. Layer 1 (the narrative of the paper) is locutionary.<sup>1</sup> Layer 2 (the argumentation layer) is illocutionary. Hypotheses are mapped from the locutionary to the illocutionary layer. Evidence to support hypotheses stems from the locutionary layer and can consist of individual textual units representing observations, results, conclusions, and background claims.

► **Definition 1. Layer 1 (Domain World):** The “semantic” layer that reflects the underlying domain information. Typical labels are *Domain Entity*, *Domain Relationship*, *Method*, or *Measurement*. The precise semantics of each label requires definition, and is probably going to differ in different annotation schemes.

► **Definition 2. Layer 2 (Rhetorical World):** The “rhetorical” layer that reflects the argumentation of the paper. This necessarily includes the author in some way, for example as holder of an opinion or observer of some fact. Typical labels are *Claim*, *Hypothesis*, *Motivation*, *Purpose*, *Observation*, *Related Work*, *Experiment*, *Model*, *Background*, or *Conclusion*. The precise semantics of each label requires a definition, and is probably going to differ in different annotation schemes. Note that the same Layer 1 units can be rearranged into different arguments by different Layers 2.

#### 5.1.4 An Example Labelset

We have in mind a modular, core annotation scheme that is domain-independent and can be extended as needed with domain specificities. Many people have worked on the components and functions of argumentation, from Aristotle [1] to Toulmin [9] and Walton et al. [10]. We do not propose a preferred set of labels as the “correct” one; we merely draw from previous work as an illustration. By adopting (some of) these labels and adding more as needed for any specific task, anyone using this framework would make their work available to others doing the same. Our labelset is drawn primarily from the following three sources.

The first source [6] includes eleven main categories, considered by the authors as the Core Scientific Concepts (CoreSC). They are listed in Table 1, including the distinction of a finer-grained classification that gives details about the properties of objects and methods mentioned in the paper.

---

<sup>1</sup> Following Austin [2], locutionary acts are our statements with their immediate and direct meanings. Illocutionary acts derive from the performance of our statements, like asserting, hypothesising, or performing. Perlocutionary acts affect the hearer indirectly after inference; for example, someone being persuaded or insulted.

■ **Table 1** Categories from the CoreSC Annotation scheme [6].

Category	Description
Hypothesis	A statement not yet confirmed rather than a factual statement
Motivation	The reasons behind an investigation
Background	Generally accepted background knowledge and previous work
Goal	A target state of the investigation where intended discoveries are made
Object-New	An entity which is a product or main theme of the investigation
Object-New-Advantage	Advantage of an object
Object-New-Disadvantage	Disadvantage of an object
Method-New	Means by which authors seek to achieve a goal of the investigation
Method-New-Advantage	Advantage of a Method
Method-New-Disadvantage	Disadvantage of a Method
Method-Old	A method mentioned pertaining to previous work
Method-Old-Advantage	Advantage of a Method
Method-Old-Disadvantage	Disadvantage of a Method
Experiment	An experimental method
Model	A statement about a theoretical model or framework
Observation	The data/phenomena recorded in an investigation
Result	Factual statements about the outputs of an investigation
Conclusion	Statements inferred from observations & results relating to research hypothesis

For the next source, we use the Argument Interchange Format (AIF) [8]. It is tailored towards modeling argumentation as a graph, but is not specific to scholarly papers. There are various types of nodes that represent ontological concepts:

- Information: The *I-node* contains the utterance (also called proposition), the minimal building block without any other rhetorical semantics.
- Anchor: The *YA-node* is about all speech acts, such as assert, hypothesize, or claim, among others. It links two *I-nodes* and represents the illocutionary forces.
- Applications of Rules of Inference or Conflict: The *RA-node* is used for connecting two *I-nodes* with inference, while the *CA-node* does the same but for conflict.
- Rephrase: The *MA-node* is for restating a proposition and includes purposes such as generalization, specification, or exemplification.
- Transition: The *TA-node* indicates the transition between two *I-nodes* and can coexist between the same two propositions parallel to another relation. They contain the dialogue relations (e.g., between a question and an answer).

Inspired by the work of Moser and Mercer [7] and Green [5], we find that the following labels are used in experimental science papers: *Premise*, *Inference*, and *Claim*. A list and taxonomy can be found in [4].

Even without formal definitions, the labelset overlaps and differences are obvious.

### 5.1.5 Definitions of Example Labels

This section lists a set of fairly generic accepted labels with definitions for each. Most exist in the above mentioned Rhetorical World (not the Domain World or the Evaluation World).

► **Definition 3. Assertion:** A simple proposition (typically a clause) that states something. In fact, every *Assertion* is a *Claim* since the implicit assumption (unless otherwise stated) is that the author believes the proposition (except perhaps for the awkward case of null hypotheses). However, we differentiate *Assertions*, which for us carry no implicit connotation that the author believes them to be true, from *Claims*, for which the author’s epistemic (truth) judgment must be given. Thus, we use a narrow interpretation of “*Claim*”. In AIF, this is called *I-Node*.

► **Definition 4. Claim:** A frame that consists of an author or a speaker (called claimer), *Claim* content (an *Assertion*), the epistemic status (which can be true, false, maybe, desired, unknown, . . .), and a set of links to support or opposition frames. In AIF, this corresponds to *YA-nodes*.

► **Definition 5. Support:** A link that connects two other frames. To be able to associate additional information with it, we reify the link and state it as a frame consisting of a *Claim*, which may even appear in another paper, and Evidence (a set of *Assertions* or *Claims*). This is called *RA-node* in AIF.

► **Definition 6. Oppose:** As *Support*, mutatis mutandis, and corresponds to *CA-node*.

► **Definition 7. Hypothesis:** A frame, which is almost identical to a *Claim* but whose epistemic status is unknown or desired. It usually appears without *Support* or *Oppose* links. In AIF, this is expressed using *YA-nodes*.

► **Definition 8. Motivation or Goal:** A frame that expresses the desired target state after an experiment has been executed consisting of a holder (a person with the goal, usually the author) and a desired state (usually a *Hypothesis*, but with its epistemic status being proved). This is included in the *YA-nodes*.

► **Definition 9. Step:** A single action (in the Domain World) performed on domain objects (from the Domain World). This corresponds to a clause and involves an actor (usually, someone from the author’s team). There are different kinds of *Step*, depending on the nature of the domain. Most experiments include a measurement (see *Assay* below), one or more observations, and one or more conclusions.

► **Definition 10. Method:** A frame of an ordered series of *Steps*.

► **Definition 11. Assay:** A frame (a more specific *Step*) consisting of an actor, a measurement (a *Method*), a metric (a measuring unit accepted in the Domain World), and a result (a number determined by the *Method* expressed in the *metric*).

► **Definition 12. Experiment:** A frame with a local *Hypothesis* (i.e., restricted to one aspect being studied), a *Method*, an *Assay*, and a result, which is a *Claim* frame whose epistemic status is proved.

► **Definition 13. Interpretation:** A frame that draws together several *Assays* into a single *Claim*. It is made up of *Experiments* (a list of *Assays*, or perhaps their *Experiments*) and conclusions, which are a set of *Claims* or *Hypotheses* with the epistemic status of proved.

► **Definition 14. Restatement:** A link that connects two other frames that have the “same” (semantic) meaning. To be able to associate additional information with it, we reify the link and state it as a frame consisting of Version 1 and Version 2 (of an *Assertion*). These are propositions, which may even appear in another paper. In AIF, this is represented by the *MA-nodes*.

► **Definition 15. (Research) Question:** A question about a proposition which is a frame consisting of a questioner (a person, usually the author) and the focus of a question (a proposition). This is included in AIF's *YA-nodes*.

► **Definition 16. Dialogue Relation:** A link that connects two other frames and expresses a dialogue function, such as a question and an answer or a full form and a summary. Typically, *Dialogue Relations* coexist between two frames that are also related using another relation in parallel. To be able to associate additional information with it, we reify the link and state it as a frame made up of the dialogue prior (which is any frame, e.g., a *Question*) and the dialogue posterior (any frame, e.g., the proposition that answers it). In AIF, this is expressed using *TA-nodes*.

### 5.1.6 Next Steps

As mentioned in the outset, we do not propose a finalized framework of frames and sets of labels. But we hope that the frames and labels listed here may serve most purposes and encourage standardization across research. It is left for future work to flesh out both the labelset(s) and the relations.

Furthermore, the development of a typology of paper genres is necessary in order to apply the framework. At the minimum, the different structures used in the genres Experiment Reports, Mathematical Proofs, and Surveys should be elaborated.

The proposed framework of genres, frames, and labelsets will be best tested by the creation of example annotation datasets.

The other three sections of this report are compatible with the framework proposed here. Section 5.3 on Evaluation develops an additional layer of labels.

### References

- 1 Aristotle (1954). *The Rhetoric and the Poetics of Aristotle* (translated by W. Rhys Roberts). Random House, New York.
- 2 Austin, J. (1975). *How to Do Things with Words*. Harvard University Press.
- 3 Burns, G. A. P. C., de Waard, A., Dasigi, P., and Hovy, E. H. (2016). Cycles of scientific investigation in discourse – machine reading methods for the primary research contributions of a paper. In Jaiswal, P., Hoehndorf, R., Arighi, C. N., and Meier, A., editors, *Proceedings of the Joint International Conference on Biological Ontology and BioCreative, Corvallis, Oregon, United States, August 1-4, 2016*, volume 1747 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- 4 Green, N. L. (2018a). Proposed method for annotation of scientific arguments in terms of semantic relations and argument schemes. In Slonim, N. and Aharonov, R., editors, *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 105–110. Association for Computational Linguistics.
- 5 Green, N. L. (2018b). Towards mining scientific discourse using argumentation schemes. *Argument Comput.*, 9(2):121–135.
- 6 Liakata, M., Saha, S., Dobnik, S., Batchelor, C. R., and Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinform.*, 28(7):991–1000.
- 7 Moser, E. and Mercer, R. E. (2020). Use of claim graphing and argumentation schemes in biomedical literature: A manual approach to analysis. In *Proceedings of the 7th Workshop on Argument Mining*, pages 88–99.
- 8 Rahwan, I. and Reed, C. (2009). The argument interchange format. In Simari, G. R. and Rahwan, I., editors, *Argumentation in Artificial Intelligence*, pages 383–402. Springer.

- 9 Toulmin, S. E. (2008). *The Uses of Argument, Updated Edition*. Cambridge University Press.
- 10 Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.

## 5.2 Cross-domain Argumentation Model for Scholarly Argumentation

*Khalid Al-Khatib (University of Groningen, NL)*


*Fengyu Cai (TU Darmstadt, DE)*

*Dayne Freitag (Artificial Intelligence Center, SRI International – Menlo Park, USA)*

*Daniel Garijo (Universidad Politécnica de Madrid, ES)*

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

*Henning Wachsmuth (Leibniz Universität Hannover, DE)*

License  Creative Commons BY 4.0 International license

© Khalid Al-Khatib, Fengyu Cai, Dayne Freitag, Daniel Garijo, Benno Stein, Henning Wachsmuth

Although all scholarly discourse shares a common set of goals that can be easily articulated – the increase of human knowledge, the achievement of consensus among scholars, etc. – it encompasses a huge variety of disciplines and objectives. It is not immediately clear that a model developed to explain argumentation in one domain, like computational linguistics, can be applied to the scholarly literature on seismology or clinical psychology. A unified view of scholarly argumentation is clearly desirable, potentially increasing the speed with which new scholarly domains can be modeled computationally. The *Domains* working group sought to investigate the feasibility of such a universal framework. Rather than approaching this question based on first principles, as in the *Foundations* working group, we adopted a comparative approach, anchoring our inquiry in a close reading of two papers from widely different domains.

### 5.2.1 Objective

The basic goal of scholarly argumentation is to add knowledge to the existing knowledge of a domain/field.<sup>2</sup> The way this knowledge is added (the kind of scholarly argumentation) follows the specific rules and traditions of the field. The definition of these specific rules and traditions is what we refer to as an argumentation type.

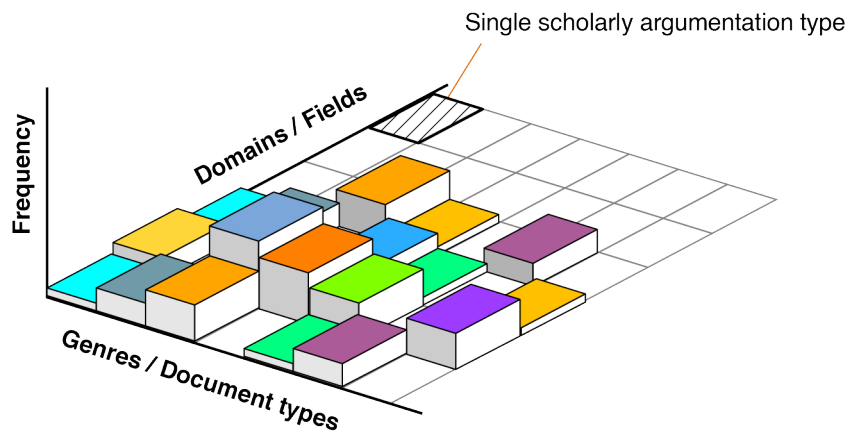
We attempt to identify different scholarly argumentation types, organized based on the domain (i.e., area of expertise) and genre (i.e., document type) of the scientific publication. Figure 2 illustrates the idea of prevalent argumentation types across different domains and different document types.

If we can acquire conceptual and empirical knowledge about the distribution illustrated in Figure 2, we will identify usage patterns across domains, and at least partially, decouple topics, domains, and argumentation types. This orthogonality can justifiably be seen as the identification of argumentation strategies.

Starting from an anecdotal study with the review of two different publications, we discuss the main gaps when annotating argumentative sentences in scientific papers.

---

<sup>2</sup> In addition to this primary purpose, we recognize that papers are also written self-expression, career reasons, or other reasons.



**Domains / Fields** (exemplary, from DFG scheme) :

- Humanities
- Sport
- Law, Economics, and Social Sciences
- Mathematics, Natural Sciences
- Human Medicine / Health Sciences
- Agriculture, Forestry and Nutritional Science, Veterinary Medicine
- Engineering
- Art, Art Theory

**Genres / Document types** (exemplary) :

- Blog posts
- Debates
- Essays
- Law texts
- News
- Political speeches
- Reviews
- Scientific articles
- Wikipedia

■ **Figure 2** Each cell corresponds to a single argumentation type, where same/similar colors hint same/similar types. There are argumentation types that are used across all domains and genres, but also domain- and genre-specific types.

### 5.2.2 Anecdotal Study

Argumentation in scientific articles may be modeled at different levels of granularity, from the macro-level discourse structure of an entire article (e.g., in terms of elements such as model, experiments, and discussion) to the micro-level argumentative structures of individual clauses, sentences, and paragraphs. As an initial basic study, all members of the breakout group annotated the sentence-level structure of the introductions of two scientific articles from different disciplines, identifying which sentences comprise the *claims* and the *premises* of the authors' arguments. Here, we considered a claim to be an assertion that the authors aim to sell as new, true, or similar, and a premise as a reason supporting either the claim directly or another premise.

In particular, we considered one paper each from two domains reflecting two different types of papers, namely, a corpus paper from computational linguistics [1] and an experiment paper from medical chemistry [2]. For each paper, we first annotated its introduction individually, and then we compared and discussed the results. Here, we report only on some noteworthy findings that we made.

First, we observed that these two texts share a similar argumentative agenda and structure, despite the wide divergence in subject matter and lexical content. As shown in Table 2, this rough similarity can be exposed by comparing key sentences drawn from different locations in a paper's introductory section. These sentences exhibit intents – which we characterized

■ **Table 2** The argumentative agendas and structures of two papers, one from computational linguistics (a corpus paper) and one from medical chemistry (an experimental paper).

Location ( <i>Audience</i> )	Computational Linguistics	Medical Chemistry
title ( <i>reviewer</i> )	“Mama Always Had a Way of Explaining Things So I Could Understand”: A Dialogue Corpus for Learning to Construct Explanations	Protein-Structure Assisted Optimization of 4,5-Dihydropyrimidine-6-Carboxamide Inhibitors of Influenza Virus Endonuclease
lead sentence ( <i>sponsor</i> )	Explaining is one of the most pervasive communicative processes in everyday life....	Influenza is an infectious disease associated with 500,000 deaths and 3–4 million severe illnesses annually....
main claim ( <i>lead researcher</i> )	We argue that a better understanding of how humans explain in dialogues is needed, so that XAI can learn to interact with humans.	Our overarching approach has been to apply structure-based design, while optimizing inhibitors... in order to proactively develop lead inhibitors that are less likely to rapidly develop clinical resistance.
proximal claim ( <i>junior researcher</i> )	In this paper, we present a first corpus for computational research on....	Here, we describe the further optimization of such a series of new endonuclease inhibitors....

in terms of putatively different audiences – that vary with their position in the discourse and are shared across the two target domains. Titles must succinctly summarize a paper’s content and, depending on the domain, may include features intended to draw interest from potential reviewers. Lead sentences typically state the overriding concern an entire field addresses, often in a language digestible by a general audience. Sentences expressing claims vary in their specificity and concreteness, ranging from concrete contribution to central insight.

As shown in the table, we distinguished between *main claims* (claims that the paper’s author presumably deemed most important) and *proximal claims* (*pro forma* claims that provide useful context). We found that the ability to distinguish these two types of claims relies substantially on domain expertise. For both domains, we observed that the introduction contains only very few real claims in the sense of assertions the authors aim to convince the reader of – about one to three depending on the annotator.

Initially, there was notable disagreement in the group, none of whom has extensive chemistry expertise, about which statements in the medical chemistry paper constituted claims and which of these was the main claim. In contrast, the group’s annotations of the computational linguistics paper showed considerable agreement. The only exception was the annotations of a group member with less background in computational linguistics. This member chose as the main claim a sentence that all other members viewed as proximal.

This result clearly established the importance of domain expertise for certain types of argumentative analysis. In particular, determining which is the *main* claim requires the reader to assess the scientific significance of a statement, an assessment that may require extensive knowledge of an area of research. However, based on our interdisciplinary discussion, we reached an agreement in most cases, even in our analysis of the chemistry paper, suggesting that the automated modeling of the argumentative structure of scientific articles is feasible, in principle. The key question is how much domain knowledge the analysis of scientific argumentation in a given discipline requires.



### 5.2.3 A New Approach: Towards Reducing Reliance on Domain Experts

Our anecdotal study and discussion suggest that having domain experts for all paper types and domains may be a costly and inefficient process. Instead, we identified a new research challenge: how to accelerate the interaction with domain experts to speed up cross-domain argumentation annotation? This research challenge spans new research questions such as:

- Can we probe experts with specific portions of text instead of having them read the whole publication?
- Can we identify a specific vocabulary and use it in customized domain-specific annotation platforms?
- Can we identify a set of questions for domain experts to help guide other users in the annotation process? Examples of these questions include identifying the section where the main claim is, which are the main sections to look at first when analyzing a paper in a particular domain, what are the main types of evidence in a publication or typical lexical cues to identify claims or evidence in a given domain.

Following our anecdotal study, we explored some of these questions with our two papers, as shown in Tables 3. We believe these are initial examples that should be expanded in order to identify a wider range of commonalities in scientific literature.

■ **Table 3** Examples of three questions that domain experts can answer to assist non-experts in the annotation process.

<i>Where is the main claim? in which section can we find it?</i>					
<b>Publication domain</b>	Introduction	Method	Experiments	Results	Discussion
NLP	X	X			
Chemistry	X			X	
<i>What are the main sections that we should look at first?</i>					
<b>Publication domain</b>	Introduction	Method	Experiments	Results	Discussion
NLP	1	2	4	3	5
Chemistry	1	5	4	2	3
<i>What are the main types of evidence in your domain?</i>					
<b>Publication domain</b>	Anecdote	Statistics	Testimony	Analogy	Figure/table
NLP		X			X
Chemistry		X		X	

### 5.2.4 Next Steps

Our working group discussed the *Introduction* section of two conference papers from computational linguistics and medical chemistry as examples to explore the discrepancy in argumentation between domains. Through manual annotation and discussion, we came to find that scientific argumentation varies among different domains noticeably. Further work may extend this analytical and comparative paradigm on scholarly argumentation to other domains, genres, and parts in the publication. After discussing the results of our anecdotal study, we believe that more research is needed to accelerate domain expert interaction for annotating argumentative sentences in different domains. Instead of asking domain experts to directly help with various task-specific works required by non-experts, their contribution would be more efficient and influential by helping summarize the universal features of argumentation for one specific genre, domain, and part. For example, experts could annotate a feasible scale of representative papers, extract lexical hints, etc.

Key questions in this area include how to structure the interaction with the domain expert for lightweight knowledge elicitation, and how to abstract, represent, and inject the features that encapsulate the knowledge required for accurate models of a given domain’s argumentation. Meanwhile, without sacrificing models’ performance, the minimum degree of domain knowledge elicitation from experts is also worth studying.

## References

- 1 Henning Wachsmuth and Milad Alshomary. “mama always had a way of explaining things so I could understand”: A dialogue corpus for learning to construct explanations. In Proceedings of the 29th International Conference on Computational Linguistics, pages 344–354, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
- 2 Diane Beylkin, Gyanendra Kumar, Wei Zhou, Jaehyeon Park, Trushar Jeevan, Chandraiah Lagiseti, Rhodri Harfoot, Richard J Webby, Stephen W White, and Thomas R Webb. Protein-structure assisted optimization of 4, 5-dihydroxypyrimidine-6-carboxamide inhibitors of influenza virus endonuclease. *Scientific reports*, 7(1):1–12, 2017.

## 5.3 Evaluation of Argument Quality

*Yufang Hou (IBM Research Europe – Dublin, IE)*

*Tobias Mayer (Technical University Darmstadt – Darmstadt, DE)*

*Domenic Rosati (scite.ai – Halifax, CA)*

*Harrisen Scells (Leipzig Universität DE)*

*Ferdinand Schlatt (Universität Halle-Wittenberg – Halle, DE)*

*Simone Teufel (University of Cambridge, UK)*

*Ryan Wang (University of Illinois Urbana-Champaign – Urbana, USA)*

**License** © Creative Commons BY 4.0 International license

© Yufang Hou, Tobias Mayer, Domenic Rosati, Harrisen Scells, Ferdinand Schlatt, Simone Teufel, Ryan Wang

### 5.3.1 Introduction

This working group focused on the evaluation of argumentation quality. We wanted to take an alternative approach to typical text synthesis evaluation. We also wanted to develop an evaluation framework that is general enough that can be applied to the numerous argumentation schemes that exist.

We settled on an approach that would determine the argumentation quality of a text through the interrogation by a Question Answering (QA) system about the argumentation within. For each genre of text, one would need to develop a series of diagnostic (and increasingly specific) questions that would reflect the quality of the reasoning or argumentation of the paper. Each genre of text would have a different set of characteristics. For example, the process for systematic reviews would ask questions about the comprehensiveness of the review, while for technical or experimental papers, questions would be related to the description of the model, and how or why it brings about an improvement in the domain.

The envisioned evaluation system would take a text to be evaluated and a question bank organised by the different genres. Depending on the genre the appropriate questions would be applied to the text and their answers rated. Note that the framework is flexible such that answers could be rated by humans and later automatically. Depending on the effectiveness of the QA system and the nature of the answers, a combination of human and automatic answer rating can be used.

■ **Table 4** Example questions that could be posed to a QA system, and descriptions of answers.

Example Question	Description of Answer
What is the typology of the paper?	E.g., empirical research, position paper, theoretical.
Which questions apply to which genres?	E.g., empirical science may be more focused on cogency while philosophy and mathematics may be more focused on reasonableness.
What are the general properties of argumentation of interest in a specific paper genre, for making questions?	
What are the domain-specific properties of argumentation of interest?	E.g., empirical science may be more focused on ...? while philosophy and mathematics may be more focused on ...?
What is the main claim of this paper?	One or two sentences from the text that should contain the claim.
What is the proof that the paper's proposed technique is better?	Two rows extracted from a table, one for the state of the art and the other for the system, containing two numbers, the system's being better.

■ **Table 5** Taxonomy of question types and their explanations.

Question Type	Explanation
Document-level assessment vs. corpora-level assessment	Intra-document vs. inter-document
Extractive evaluation vs. reasoned evaluation	Extractive: questions that can be answered through passages in the text; Reasoned: questions that must be answered through reasoning
Content vs. Form	Content: how well are the arguments presented? Form: how well are the arguments structured?

### 5.3.2 Evaluation Framework

We developed several initial question banks for academic papers. In Table 4 we provide a sample of what we believe to be the kinds of questions that should be asked. However, how would one judge the answers? Ideally questions would be simple and easy to judge automatically. In reality the argument is nuanced and complex. Therefore, assessing the quality of answers is likely to be a human task, though as answers become more formalised (even simplified to just yes or no questions) automated assessment becomes more feasible.

We note that it seems natural that the assessor might want to record caveats, concerns, or other thoughts. We allow the assessor to include such comments as motivation for why they assign the score they do.

To guide the development of questions, we also devised a taxonomy of question types. Table 5 contains our initial taxonomy of question types. However, in addition to question types, it is also necessary to define how answer to questions will be evaluated.

Thus finally, we devised a hierarchy of evaluation. Each level corresponds to a different interrogation method for probing argumentation quality.

1. First level of evaluation: model evaluation as retrieval
  - Input: Open-ended questions
  - Output: “retrieval unit” i.e., sentence/snippet/etc.
2. Second level of evaluation: model evaluation as a checklist
  - Input: Yes/No questions
  - Output: Yes/No

3. Third level of evaluation: multiple-choice QA
  - Input: multiple-choice questions
  - Output: selection of one answer from set of answers

### 5.3.3 Next Steps

We have already begun the development of a tool for the community to perform offline evaluation of argumentation quality. We are developing the tool as an open source project, and is available at [https://github.com/hscells/arg\\_eval](https://github.com/hscells/arg_eval). We plan to continue to develop this tool to support the various question types and levels of evaluation. Once we have laid the groundwork with a proper evaluation tool and expanded upon the framework proposed here, the next logical step is the development of a QA system. The first version of the QA system will focus on a small subset of the possible question types and perhaps only one level of evaluation. This will demonstrate the viability of a QA system to evaluation argumentation quality and will set a clear direction for further expansion of the QA system.

## 5.4 Scholarly Argumentation as a Community Dialogue

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

*Andreas Vlachos (University of Cambridge, UK)*

*Davide Ceolin (Centrum Wiskunde & Informatica, Amsterdam, NL)*

*Milad Alshomary (Leibniz Universität Hannover, Germany))*

*Nils Dycke (TU Darmstadt, DE)*

*Sukannya Purkayastha (TU Darmstadt, DE)*

*Iryna Gurevych (TU Darmstadt, DE)*

*Anne Lauscher (Universität Hamburg, DE)*

*Tilman Beck (TU Darmstadt, DE)*

License © Creative Commons BY 4.0 International license

© Wolf-Tilo Balke, Andreas Vlachos, Davide Ceolin, Milad Alshomary, Nils Dycke, Sukannya Purkayastha, Iryna Gurevych, Anne Lauscher, Tilman Beck

### 5.4.1 Motivation

In science, peer reviewing is the deliberation process where members of a scientific community with diverse levels of experience decide if a scholarly work provides a valuable, scientific contribution [1, 2]. In the process, the actors of the community (i.e. authors, reviewers, meta-reviewers, and possibly others, e.g., chairs) exchange arguments about the strengths and weaknesses of a particular scientific contribution within multiple, direct and indirect dialogues (review, rebuttal, decision-making).

Usually, the decision-making process begins with the reviewers writing their reviews and optionally the authors responding to the reviews (i.e. rebuttal). Here, the **meta-reviewer has to arrive at a decision about the promotion of acceptance of the paper**. This process is mainly about weighing the arguments raised by the reviewers and happens under time constraints. To provide more efficient and effective access to (a) the content of the paper, and (b) the many arguments raised by the individual reviewers, we envision an intelligent dialogue system which answers questions of the meta-reviewer.

From an NLP perspective, this is more challenging than other domain-specific task-oriented dialog system scenarios [9], as the meta-reviewer’s needs underpinning these questions can vary from information retrieval and exploration (e.g. “What datasets did the authors use?”)

to combining information from multiple sources (e.g. “According to the reviews, what are the main weaknesses of the paper?”) and summarization tasks (e.g. “Please briefly summarize the paper?”).

The goal of this breakout group was to evaluate the feasibility of collecting (training) data for such a system and to refine the task definition along the way. Having such a dataset could provide an interesting basis for studying both various facets of argumentation, like quality and convincingness of reviews or implicit ranking of the value system employed by the meta-reviewer. Furthermore, different dialogue strategies can be analyzed, like the way of gathering information in order to come to a decision.

#### 5.4.2 Summary and Conclusions

The breakout group defined the goal of the sessions as **formulating the decision-making process for a scholarly paper as a dialogue** and conducted a first annotation round using an Oxford-style inspired debate format. Two groups (debaters and judges) were involved in the decision-making process. Given a paper and its reviews, the debaters discussed the pros and cons of the paper and a decision was formed by the judges. To study the relation between arguments extracted from the reviews or the underlying paper, all turns required explicit grounding in the respective documents. For instance, an argument in favor of acceptance should be substantiated by the review passage (*As reviewer 1 says . . .*) from which it was derived.

After reviewing the annotation process, it became clear that the task needs to be better aligned with the actual review process of the respective research discipline (in our example: Natural Language Processing) and in such a way that the data collected will be useful for a real-world system. There was a consensus that the debate format is obstructive as it forces dialogue partners to defend a position which might be different from their own. Additionally, the coarse granularity of groundings in natural dialogue – i.e. referring to the entire document instead of sentences or paragraphs – limited the study of the relations between argumentative units in the reviews and papers.

We revised the system’s purpose as a decision-making support system for the meta-reviewer after reviews (and rebuttals) are collected. Therefore, the dialogue involves two parties (meta-reviewer, intelligent support system) with the meta-reviewer questioning the system to inform their final decision, and the system as an oracle with knowledge of the paper, the reviews and optionally other related work. To resemble a real-world situation, a time limit is imposed on the meta-reviewer which enforces limited exposure to the reviews and paper. It is important to note that such a system will be most beneficial for papers where the decision is difficult (i.e. so-called *borderline papers*). Further, the system will support in weighing the reviews as there exist different levels of reviewing expertise.

Finally, we conducted another round of data collection by pairing the senior members of the group (meta-reviewers) with the junior members (imitating the dialogue system). The junior members prepared themselves by reading the papers and reviews in detail. Before the dialogue, meta-reviewers had five minutes to study the reviews. At the end of the dialogue, the meta-reviewer had to make a statement about the acceptance or rejection of the paper. We collected 16 dialogues (in English) about 4 papers involving 4 meta-reviewers and 4 system agents. The conversations were transcribed using the OpenAI Whisper [10] model which is known to have good transcription quality. However, manual post-processing was necessary as the model output is not separated based on the speaker.

In summary, we formalized the idea of a decision-support system for meta-reviewers during peer reviewing as a dialogue system. We designed and evaluated a protocol to collect dialogue data for such a system. As a result we created a dataset of 16 high-quality question & answer dialogues between a meta-reviewer and a system agent which is knowledgeable about the paper and reviews.

### 5.4.3 Challenges And Next Steps

There exist several open questions about the future course of this project. State-of-the-art NLP models require a certain amount of data for training. However, as the data collection procedure requires the participation of expert-level reviewers (i.e. meta-reviewers), it cannot be scaled easily. One way could be to align the discussion format between reviewers and meta-reviewer during peer-reviewing with the dialogue format proposed in this group. Similar to the first pilot annotation, the study of grounding of the assistants' turns in the review texts is one important future step towards modeling such alignments. This step might be facilitated by adding structured annotations to the reviews, indicating, for example, the targets of the comments (comments regarding specific parts of the paper, the experimental settings, etc.) or the severity of the issues raised by the reviewers [3].

Another question is whether the data collection procedure can be generalized to different use-cases. Here, the first step should be to separate task-specific components (e.g. meta-reviewer role) from the more general aspects. Further, an additional layer of annotation would help specify general and domain-specific dialogue acts. The usability of different dialogue argumentation schemata [4] needs to be assessed. Annotation can be conducted using off-the-shelf tools, like INCEPTION [5] or PEER<sup>3</sup>.

A crucial issue is the evaluation of the success of the conversation. As stated above, the goal of the system is to inform the meta-reviewer's final decision. This is rather difficult to quantify and can be biased by other influencing factors, e.g. low-quality reviews. Conducting user studies is a possible direction but it is costly and time-consuming. Another approach could be to assess whether individual questions have been answered satisfactorily by the system rather than directly evaluating the overall conversation. While we successfully transcribed the audio data collected in this group, we recommend data collection via text-based input methods [6] to overcome the need for post-hoc manual speaker identification and enabling data collection in an online setup. Also, we point out that neither the rebuttals nor the official meta-reviews are included in the peer review dataset [7] due to the complicated licensing situation with peer-reviewing data [8], but they would be another useful resource.

### References

- 1 Tom Jefferson and Elizabeth Wager Frank Davidoff, *Measuring the quality of editorial peer review*. JAMA, American Medical Association, pp.2786–2790, 2002.
- 2 Aliaksandr Birukou and Joseph R. Wakeling and Claudio Bartolini and Fabio Casati and Maurizio Marchese and Katsiaryna Mirylenka and Nardine Osman and Azzurra Ragone and Carles Sierra and Aalam Wassef, *Alternatives to Peer Review: Novel Approaches for Research Evaluation*. Frontiers Computational Neuroscience, p.56, 2011.
- 3 Cristina-Iulia Bucur and Tobias Kuhn and Davide Ceolin, *Peer Reviewing Revisited: Assessing Research with Interlinked Semantic Comments*. K-CAP, pp.179–187, 2019.

---

<sup>3</sup> The PEER tool is an annotation tool designed for the domain of scholarly articles permitting span annotations and commenting directly inside an article's PDF. This tool is under development at the UKP Lab, Technical University of Darmstadt; please refer to <https://intertext.ukp-lab.de/> for updates on its release

- 4 Georgi Karadzhov and Tom Stafford and Andreas Vlachos, *DeliData: A dataset for deliberation in multi-party problem solving*. arXiv preprint 2108.05271, 2021.
- 5 Jan-Christoph Klie and Michael Bugert and Beto Boudlosa and Richard Eckart de Castilho and Iryna Gurevych, *The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation*. COLING, pp.5–9, 2018.
- 6 Lorenz Stangier and Ji-Ung Lee and Yuxi Wang and Marvin Müller and Nicholas Frick and Joachim Metternich and Iryna Gurevych, *TexPrax: A Messaging Application for Ethical, Real-time Data Collection and Annotation*. ACL/IJCNLP, pp.9–16, 2022.
- 7 Nils Dycke and Ilia Kuznetsov and Iryna Gurevych, *NLPeer: A Unified Resource for the Computational Study of Peer Review*. arXiv preprint 2211.06651. 2022.
- 8 Nils Dycke and Ilia Kuznetsov and Iryna Gurevych, *Yes-Yes-Yes: Donation-based Peer Reviewing Data Collection for ACL Rolling Review and Beyond*. arXiv preprint 2201.11443, 2022.
- 9 Chia-Chien Hung and Anne Lauscher and Simone Paolo Ponzetto and Goran Glavas, *DS-TOD: Efficient Domain Specialization for Task-Oriented Dialog*. ACL, pp.891–904, 2022.
- 10 Alec Radford and Jong Wook Kim and Tao Xu and Greg Brockman and Christine McLeavey and Ilya Sutskever, *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv preprint 2212.04356. 2022.

## Participants

- Khalid Al-Khatib  
University of Groningen, NL
- Milad Alshomary  
Leibniz Universität  
Hannover, DE
- Wolf-Tilo Balke  
TU Braunschweig, DE
- Tilman Beck  
TU Darmstadt, DE
- Elena Cabrio  
Université Côte d'Azur –  
Sophia Antipolis, FR
- Fengyu Cai  
TU Darmstadt, DE
- Davide Ceolin  
CWI – Amsterdam, NL
- Anita de Waard  
Elsevier – Jericho, US
- Nils Dycke  
TU Darmstadt, DE
- Dayne Freitag  
SRI – Menlo Park, US
- Daniel Garijo  
Polytechnic University of  
Madrid, ES
- Iryna Gurevych  
TU Darmstadt, DE
- Graeme Hirst  
University of Toronto, CA
- Yufang Hou  
IBM Research – Dublin, IE
- Eduard H. Hovy  
Carnegie Mellon University,  
Pittsburgh, US & University of  
Melbourne, AU
- Anne Lauscher  
Universität Hamburg, DE
- Maria Liakata  
Queen Mary University of  
London, GB
- Tobias Mayer  
TU Darmstadt, DE
- Robert Mercer  
University of Western Ontario –  
London, CA
- Smaranda Muresan  
Columbia University –  
New York, US
- Preslav Nakov  
MBZUAI – Abu Dhabi, AE
- Sukannya Purkayastha  
TU Darmstadt, DE
- Chris Reed  
University of Dundee, GB
- Domenic Rosati  
scite – Halifax, CA
- Florian Ruosch  
Universität Zürich, CH
- Harrison Scells  
Universität Leipzig, DE
- Ferdinand Schlatt  
Universität Halle-  
Wittenberg, DE
- Benno Stein  
Bauhaus-Universität Weimar, DE
- Simone Teufel  
University of Cambridge, GB
- Serena Villata  
Université Côte d'Azur –  
Sophia Antipolis, FR
- Andreas Vlachos  
University of Cambridge, GB
- Henning Wachsmuth  
Universität Paderborn, DE
- Ryan Wang  
University of Illinois  
Urbana-Champaign, USA

