Connecting the Dots: Toward the Argument Web of Science

Florian Ruosch

Department of Informatics, University of Zurich, Zürich, Switzerland

Abstract

The *Argument Web of Science* is the vision of a knowledge graph representing arguments from scientific publications that are interlinked and dereferenceable by URIs. In this dissertation work, we pave the way toward the Argument Web of Science by exploring available tools from *Argument Mining* and by identifying current gaps. We develop a unifying benchmark to assess the state of the art for the automated extraction of arguments from natural language text. Then, we improve upon these results by deploying recombination and ensemble methods in the second step. Finally, we investigate cross-document argumentative relations to form a multi-document *Argument Web*.

1 Introduction

Nowadays, it is seemingly impossible to keep up with the enormous amount of new scientific publications as their volume is growing exponentially (Bornmann et al., 2021). Automated methods to extract a structured representation from natural language texts are needed to handle this flood of information. Such cutting-edge approaches are currently being developed at the intersection of Artificial Intelligence (AI) and Natural Language Processing (NLP). Meanwhile, the *Semantic Web* (Berners-Lee et al., 2001) has been dealing with the encoding of semantics and knowledge graphs for well over 20 years.

Scientific publications inherently have an argumentative nature (Walton and Zhang, 2013), as they tend to be persuasive monologues (Reed, 1998). Hence, argumentation is a suitable representation of the knowledge contained in scholarly documents. A graph of URI-dereferenceable and interlinked arguments is called an *Argument Web* (Rahwan et al., 2007). Accordingly, we dub our goal the *Argument Web of Science* (Ruosch et al., 2024).

Such a knowledge graph could help move scientific communication from its currently document-centric perspective to a knowledge-based view (Auer et al., 2018). It can be used to organize and retrieve scholarly information in a more structured way. Providing updates in real-time would facilitate the process of looking up the latest related work and make it more accessible across domains.

But considering the massive amount of scientific publications, manual annotation of such a graph is not feasible, and we must turn to automated extraction instead (Reed et al., 2017). Thereby, we trade off the accuracy of the human annotator for the scalability of machine-assisted methods. Annotating arguments, especially in scholarly texts, is a laborious and complicated task, requiring a lot of time and domain knowledge (Accuosto et al., 2021).

Argument Mining (AM) (Budzynska and Villata, 2016; Lawrence and Reed, 2020) is a research field on identifying argumentative structures in natural language texts. Its automated tools can extract argumentative components and predict their relations (Stab et al., 2014). These then constitute an Argument Web, to which we can add more papers, creating one interconnected knowledge graph. But how far away from implementing this are we? Hence, this dissertation work considers the following main research question:

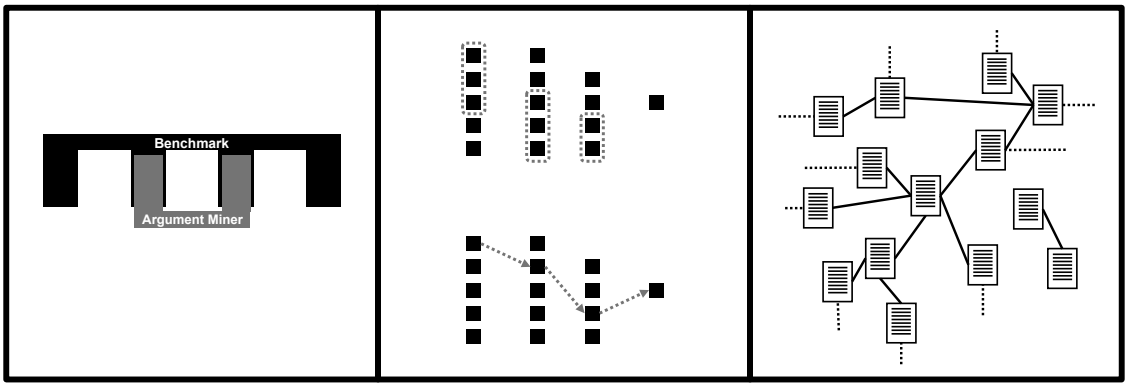

Figure 1: Overview of the three work packages in this Ph.D. project.

How can we apply existing tools to progress toward the Argument Web of Science, and what are we missing?

Figure 1 shows an overview of the three parts of the overall project, each tackling an identified gap on the way to the Argument Web of Science. In the first phase, we developed and implemented a benchmark for AM to enable the unifying and fair comparison of results to find the most accurate system. For the second step, we set out to improve the previously evaluated state of the art by leveraging ensemble methods and recombination of systems. The third and final stage involved investigating the differences between single- and multi-document argumentative relations, one of the steps to create an interconnected Argument Web.

2 Benchmarking Argument Mining

Since we rely on automated methods to extract arguments from scientific papers, we first must evaluate how well these perform. When comparing annotations from two entities (e.g., evaluating a system output against a human-annotated ground truth), we face several challenges (Lippi and Torroni, 2016). The data format and the argument representation might differ, hampering a direct comparison. Also, the annotations can be of varying granularity, such as sentences, fragments, or token-level.

To combat these issues, we introduce **BAM**, a **B**enchmark for **A**rgument **M**ining (Ruosch et al., 2022). The leftmost part of Figure 1 depicts the architecture of BAM: four pillars and the integration of the evaluated argument miner. For the data, we rely on Sci-Arg (Lauscher et al., 2018), a corpus of 40 fully argument-annotated papers from the domain of computer graphics. Furthermore, we introduce a mapping of the argument representation between the ground truth and the systems in the benchmark. By employing `subclass`- and `equivalence`-relations, we reduce to the claim/premise-model (Walton, 2009) with attack and support to enable a unifying view.

The first pillar represents the preprocessing step, where the system-specific mapping allows the tailoring of the benchmark data to fit the needs of the argument miner. In the second phase, the AM system is trained, if applicable, before it is used to annotate the test set in the third stage. The fourth and final pillar is the evaluation based on Lippi and Torroni's AM pipeline (cf. Figure 2): *sentence classification*, *boundary detection*, *component identification*, *structure prediction*.

The used metrics are anchored in NLP literature. Since the sentence classification is binary (argumentative, non-argumentative), we employ the F1-Score (van Rijsbergen, 1979). The component boundary detection is a segmentation task, and, hence,

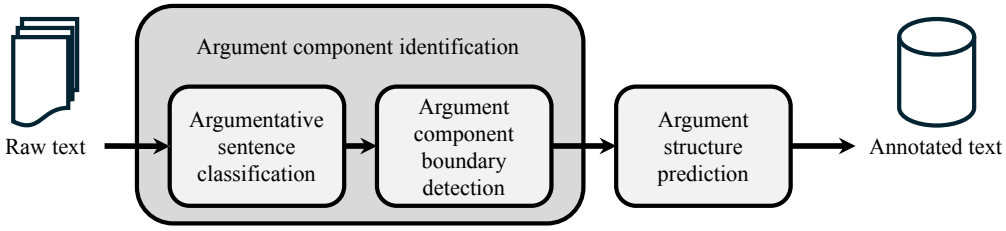

Figure 2: Argument Mining pipeline adapted from Lippi and Torroni (2016).

we apply a boundary similarity measure (Fournier, 2013). As the component identification has been pointed out to be similar to *Named Entity Recognition* (Al Khatib et al., 2021), we rely on nervaluate (Segura-Bedmar et al., 2013) and treat the argumentative components as named entities. Finally, the argumentative relation prediction is evaluated by assessing if the correct triples (subject, relation, object) have been retrieved. Hence, we utilize the F1-Score again.

We showcase the benchmark with five argument miners and demonstrate its ability to produce comparable results. Furthermore, a formal evaluation of the proposed concepts in BAM is currently forthcoming to validate its scientific soundness.

3 Improving Argument Mining

In the second part of the dissertation work, we developed **DREAM**, a framework for the **D**eployment of **R**ecombination and **E**nsembles for **A**rgument **M**ining (Ruosch et al., 2023). We aim to combat the holistic and black-box view of the AM pipeline (Lawrence and Reed, 2020) and advance the previously evaluated state of the art in AM.

To this end, we propose three approaches, partially depicted in the center section of Figure 1. First, we apply ensemble methods (Opitz and Maclin, 1999) to combine sets of two or more argument miners to improve in single tasks in the four-stage AM pipeline (see Figure 2). This is shown in the top part of Figure 1, where the squares represent the modules of AM systems for the tasks, and the dotted lines indicate the chosen ensembles. Second, we explore recombination throughout the pipeline by feeding the intermediate results of one system into the input of another to augment its output. This is displayed in the bottom part, where the dotted arrows signify the "chosen path" through different systems for the pipeline.

In our experiments, we show that applying ensemble methods (voting, stacking, bagging) can improve accuracy for single tasks. Furthermore, the same holds for recombination, where using the most accurate system's intermediate results can lead to higher accuracy in subsequent tasks for other systems.

Finally, we combine these two approaches by allowing the intermediate results for the recombination also to having been produced by ensembles. In some cases, this leads to improved accuracy. Hence, we have demonstrated the use of DREAM to advance the state of the art in AM.

4 Extending Argument Mining

The third and final step is shown in the rightmost part of Figure 1: creating a web of documents. In **MIDAS**, **M**ining **I**nter-**D**ocument **A**rguments in **S**cientific papers, we investigate cross-document argumentative connections. The previously employed Sci-Arg dataset (Lauscher et al., 2018) contains 40 papers with argumentative components and relations annotated, but each item is self-contained with no links to other documents. In this stage, we aim to address this to enable a proper Argument Web of Science.

First, we identify inter-document argumentative relations and distinguish them from intra-document links to augment the dataset. We do this by relying on already annotated components that are references to other documents: citations. The set of relations

is divided into intra-document and inter-document, allowing for a clear distinction between the two. Furthermore, we resolve the citations using the bibliography and add these papers, augmenting Sci-Arg and extending it to over 800.

Then, we evaluate three baseline approaches for argumentative relation prediction on the newly created dataset. The first method is rule-based, relying on the presence of discourse indicators that indicate attack (e.g., "however") or support (e.g., "because"). The second technique leverages the relation prediction module of a transformer-based argument miner (Mayer et al., 2020). Third, we use an off-the-shelf LLM—namely, Mistral (Jiang et al., 2023))—and prompt it in zero- and few-shot settings.

Surprisingly, the naive rule-based approach outperforms the sophisticated neural methods. Also, we find significant differences in the results when comparing the intra- and inter-document settings. This indicates that a distinction is necessary, and the two cases require dedicated approaches to predict argumentative relations.

5 Conclusions

In this dissertation work, we push toward the Argument Web (Rahwan et al., 2007) of Science (Ruosch et al., 2024): the vision of a knowledge graph comprised of interlinked and dereferenceable arguments from scientific publications. We identified the available tools and gaps that need to be filled to use them effectively to produce such a structured representation of the scholarly discourse.

First, we designed and implemented a benchmark to evaluate AM systems in a unified way to enable a fair and homogenous comparison of their results. This also allows one to find the best-suited tool for a given AM task and, more importantly, the most accurate argument miner. **BAM** (Ruosch et al., 2022) is based on a four-stage pipeline (Lippi and Torroni, 2016): *sentence classification*, *boundary detection*, *component identification*, and *relation prediction*. For each task, we found and assigned a metric to quantify the accuracy of any given system on a scale of zero to one, with bigger signifying better. To showcase BAM, we evaluated a range of current AM systems.

Subsequently, we developed a way to advance the state of the art in AM without figuratively "reinventing the wheel." With the **DREAM** (Ruosch et al., 2023) framework, we leverage recombination and ensemble methods. This allowed for improvements in the results observed in the BAM showcase, proving the efficacy of our approach. Furthermore, by splitting the previously end-to-end systems according to the four-stage pipeline, we also partially deconstructed the holistic black-box approaches and obtained their intermediate results.

Finally, we addressed the challenge of multi-document argumentative relation prediction for scientific papers. Identifying such links between documents is central to moving from independent, unconnected Argument Webs to one large knowledge graph. Hence, we extended the existing dataset (Lauscher et al., 2018) from 40 items to over 800 and explicitly annotated relations across document boundaries. By then applying three baseline approaches (discourse-indicator-based, an argument miner, and an out-of-the-box LLM), we also found statistically significant differences between the intra- and inter-document settings. This indicates that we need to distinguish between the two, and link prediction methods cannot treat them equivalently; instead, we need specialized approaches.

Overall, we advocated for more standardization to facilitate the data exchange and to harmonize the AM pipeline (Lawrence and Reed, 2020). We hope to have provided valuable tools to progress toward the Argument Web of Science. The next step would be to bootstrap a larger knowledge graph by curating a well-annotated dataset of papers.

Acknowledgements

The author thanks Abraham Bernstein and Cristina Sarasua for their guidance, insights, and support.

This work was partially funded by the Swiss National Science Foundation (SNSF) through project *CrowdAlytics* (Grant Number 184994).

References

Accuosto, P., Neves, M., and Saggion, H. (2021). Argumentation mining in scientific literature: From computational linguistics to biomedicine. In Frommholz, I., Mayr, P., Cabanac, G., and Verberne, S., editors, *Proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 43rd European Conference on Information Retrieval (ECIR 2021), Lucca, Italy (online only), April 1st, 2021*, volume 2847 of *CEUR Workshop Proceedings*, pages 20–36. CEUR-WS.org.

Al Khatib, K., Ghosal, T., Hou, Y., de Waard, A., and Freitag, D. (2021). Argument mining for scholarly document processing: Taking stock and looking ahead. In Beltagy, I., Cohan, A., Feigenblat, G., Freitag, D., Ghosal, T., Hall, K., Herrmannova, D., Knoth, P., Lo, K., Mayr, P., Patton, R. M., Shmueli-Scheuer, M., de Waard, A., Wang, K., and Wang, L. L., editors, *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65, Online. Association for Computational Linguistics.

Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., and Vidal, M. E. (2018). Towards a knowledge graph for science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, WIMS '18, New York, NY, USA. Association for Computing Machinery.

Berners-Lee, T., Hendler, J., Lassila, and Ora (2001). The Semantic Web. *Scientific American*, 284(5):34–43.

Bornmann, L., Haunschild, R., and Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):224.

Budzynska, K. and Villata, S. (2016). Argument mining. *IEEE Intell. Informatics Bull.*, 17(1):1–6.

Fournier, C. (2013). Evaluating text segmentation using boundary edit distance. In Schuetze, H., Fung, P., and Poesio, M., editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712, Sofia, Bulgaria. Association for Computational Linguistics.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b. *CoRR*, abs/2310.06825.

Lauscher, A., Glavaš, G., and Ponzetto, S. P. (2018). An argument-annotated corpus of scientific publications. In Slonim, N. and Aharonov, R., editors, *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

Lawrence, J. and Reed, C. (2020). Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10:1–10:25.

Mayer, T., Cabrio, E., and Villata, S. (2020). Transformer-based argument mining for healthcare applications. In Giacomo, G. D., Catalá, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., and Lang, J., editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press.

Opitz, D. W. and Maclin, R. (1999). Popular ensemble methods: An empirical study. *J. Artif. Intell. Res.*, 11:169–198.

Rahwan, I., Zablith, F., and Reed, C. (2007). Laying the foundations for a world wide argument web. *Artif. Intell.*, 171(10-15):897–921.

Reed, C. (1998). Is it a Monologue, a Dialogue or a turn in a Dialogue? In *Proceedings of the 4th International Conference on Argumentation (ISSA98)*.

Reed, C., Budzynska, K., Duthie, R., Janier, M., Konat, B., Lawrence, J., Pease, A., and Snaith, M. (2017). The argument web: an online ecosystem of tools, systems and services for argumentation. *Philosophy & Technology*, 30(2):137–160.

Ruosch, F., Sarasua, C., and Bernstein, A. (2022). BAM: Benchmarking Argument Mining on Scientific Documents. In Veyseh, A. P. B., Dernoncourt, F., Nguyen, T. H., Chang, W., and Lai, V. D., editors, *Proceedings of the Workshop on Scientific Document Understanding co-located with 36th AAAI Conference on Artificial Intelligence, SDU@AAAI 2022, Virtual Event, March 1, 2022*, volume 3164 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ruosch, F., Sarasua, C., and Bernstein, A. (2023). DREAM: Deployment of Recombination and Ensembles in Argument Mining. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5277–5290, Singapore. Association for Computational Linguistics.

Ruosch, F., Sarasua, C., Reed, C., and Bernstein, A. (2024). Toward the Argument Web of Science. In *Computational Models of Argument - Proceedings of COMMA 2024, Hagen, Germany, 18-20 September 2024*, volume 388 of *Frontiers in Artificial Intelligence and Applications*, pages 365–366. IOS Press.

Segura-Bedmar, I., Martínez, P., and Herrero-Zazo, M. (2013). SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In Manandhar, S. and Yuret, D., editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Stab, C., Kirschner, C., Eckle-Kohler, J., and Gurevych, I. (2014). Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In Cabrio, E., Villata, S., and Wyner, A. Z., editors, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014*, volume 1341 of *CEUR Workshop Proceedings*. CEUR-WS.org.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth.

Walton, D. (2009). Argumentation theory: A very short introduction. In Simari, G. R. and Rahwan, I., editors, *Argumentation in Artificial Intelligence*, pages 1–22. Springer.

Walton, D. and Zhang, N. (2013). The epistemology of scientific evidence. *Artif. Intell. Law*, 21(2):173–219.