Computational Models of Argument
C. Reed et al. (Eds.)

doi:10.3233/FAIA240341

A Unified Benchmark for Argument Mining

Florian RUOSCH [a,1], John LAWRENCE [b], Cristina SARASUA [a] and Abraham BERNSTEIN [a]
[a] *Department of Informatics, University of Zurich, Switzerland*
[b] *Centre for Argument Technology, University of Dundee, UK*

Keywords. Argument Mining, Benchmarking, Evaluation

1. Introduction

A unifying way to evaluate Argument Mining (AM) systems is desirable but remains challenging [1]. Still, in order to warrant comparability of the results of AM tools, such a unification for the evaluation method is necessary. The advent of the transformer architecture [2] and other advances in NLP have brought improvements in the field of AM, yet we still lack a homogeneous evaluation routine due to a variety of problems. For instance, argument miners take as inputs and produce outputs of different levels of granularity, which precludes their comparison. Furthermore, various ways to represent an argument in data are in use, originating in philosophy and diverse worldviews. Finally, there is a wide spectrum of measures applied to AM, meaning that the comparison of systems from different authors becomes cumbersome or, in some instances, impossible.

2. BAM: Benchmarking Argument Mining

To address the challenges stated above, we developed BAM [3], a **B**enchmark for **AM**, based on the four-stage AM pipeline [1]: *sentence classification*, *boundary detection*, *component identification*, and *relation prediction*. Hereby, Argument Mining is broken down into four sequential tasks. First, sentences are classified as argumentative or non-argumentative. Then, the boundaries of argumentative spans (i.e., components) are detected by segmenting the text. In the third step, the class of these components is identified according to an argument model defined beforehand. To unify the results in the benchmark and make them comparable, we use a mapping to simplify any representation of an argument to the *claim/premise* model. In the last stage, the relations between the components are predicted from a pre-defined set; in our case *attacks*, *supports*, and *not-related*. This also enables a simplification using a mapping for the relations. As the dataset, the initial implementation uses Sci-Arg [4]: 40 fully argument-annotated papers from the domain of computer graphics.

[1] Corresponding Author: Florian Ruosch, ruosch@ifi.uzh.ch.

In addition to the original publication, where we showcased BAM on five argument miners, we extended it to allow for finding statistically significant differences in the results when comparing sets of AM systems. The implementation and documentation are available in the online code repository.[2]

3. The Impact of BAM

The recent rapid growth in AM shows that there is an increasing demand for the automated extraction of deeper meaning from the vast amounts of data that we currently produce. Argument Mining techniques continually improve performance on extracting details of the argumentative structure expressed within a piece of text, focusing on different levels of argumentative complexity as the domain and task require. However, there is currently no agreed-upon way of comparing results where techniques perform at different levels of granularity or against data with partially conflicting notions of argument. BAM provides a framework for comparison of the four key tasks in the AM pipeline, offering a clearer picture of how techniques compare and how different techniques may be best combined to produce an overall pipeline with the best possible results. In this demo, we aim to showcase the value of BAM comparisons, collect feedback from the community to guarantee the framework's usefulness, and promote the benchmarking culture in the field.

Acknowledgments

This research was partially funded by the Swiss National Science Foundation (SNSF) through projects "CrowdAlytics" (contract 184994) and "Digital Deliberative Democracy" (contract 205975).

References

[1] Lippi M, Torroni P. Argumentation Mining: State of the Art and Emerging Trends. ACM Trans Internet Techn. 2016;16(2):10:1-10:25. Available from: https://doi.org/10.1145/2850417.

[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, et al., editors. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA; 2017. p. 5998-6008. Available from: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[3] Ruosch F, Sarasua C, Bernstein A. BAM: Benchmarking Argument Mining on Scientific Documents. In: Veyseh APB, Dernoncourt F, Nguyen TH, Chang W, Lai VD, 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. vol. 3164 of CEUR Workshop Proceedings. CEUR-WS.org; 2022. Available from: https://ceur-ws.org/Vol-3164/paper5.pdf.

[4] Lauscher A, Glavaš G, Ponzetto SP. An Argument-Annotated Corpus of Scientific Publications. In: Slonim N, Aharonov R, editors. Proceedings of the 5th Workshop on Argument Mining. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 40-6. Available from: https://aclanthology.org/W18-5206.

[2] https://gitlab.ifi.uzh.ch/DDIS-Public/BAM