Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Demonstrations Track

ACTA 2.0: A Modular Architecture for Multi-Layer Argumentative Analysis of Clinical Trials

Benjamin Molinet[1] , **Santiago Marro**[1] , **Elena Cabrio**[1] , **Serena Villata**[1*] and **Tobias Mayer**[2]
[1]Université Côte d'Azur, Inria, CNRS, I3S, France
[2]Technische Universität Darmstadt, Germany
{benjamin.molinet, santiago.marro, elena.cabrio, serena.villata}@univ-cotedazur.fr,
tmayer@ukp.informatik.tu-darmstadt.de

Abstract

Evidence-based medicine aims at making decisions about the care of individual patients based on the explicit use of the best available evidence in the patient clinical history and the medical literature results. Argumentation represents a natural way of addressing this task by *(i)* identifying evidence and claims in text, and *(ii)* reasoning upon the extracted arguments and their relations to make a decision. ACTA 2.0 is an automated tool which relies on Argument Mining methods to analyse the abstracts of clinical trials to extract argument components and relations to support evidence-based clinical decision making. ACTA 2.0 allows also for the identification of PICO (Patient, Intervention, Comparison, Outcome) elements, and the analysis of the effects of an intervention on the outcomes of the study. A REST API is also provided to exploit the tool's functionalities.

1 Introduction

Argumentation is the process by which arguments are constructed, compared and evaluated to establish whether any of them is warranted. Argumentation is an effective approach for solving various theoretical and practical problems [Atkinson *et al.*, 2017], like explaining and justifying decision making outcomes. This is particularly important in evidence-based medicine, where high quality evidence is combined with the individual clinical experience of the practitioner with respect to the patient's values, to achieve the best possible outcome [Sackett and Rosenberg, 1995].

In this paper, we present a new version of ACTA [Mayer *et al.*, 2019], a web tool designed to assist clinicians in analyzing clinical trials, empowering them with the ability to retrieve the main arguments (i.e., the claims and evidence together with their relations of support or attack) contained in the text. ACTA 2.0[1] proposes several new functionalities compared to the previous version. First, it exploits novel state-of-the-art neural models for the analysis of clinical trials, going beyond the link prediction task by labelling argument relations (i.e., to label as *attack* or *support* the relations holding between the identified argument components). Second, it goes beyond PICO element identification by allowing for the automatic analysis of the effect of an intervention on the observed outcome parameters [Mayer *et al.*, 2021]. Finally, a refactoring of the ACTA architecture allows now for a modular solution, such that, through a REST API, the different ACTA 2.0 modules are made available to be integrated by external systems. The tool is currently used in the context of some collaborations with hospitals in Nice (France) and for the analysis of the publications of the French National Institute of Health and Medical Research (INSERM).

To the best of our knowledge, ACTA 2.0 is the only automated tool which allows for a deep analysis of clinical text from the argumentative point of view to support evidence-based medicine. Few systems tackle similar tasks, like EVIDENCEMINER [Wang *et al.*, 2020] (which, given a natural language query, automatically retrieves sentence-level textual evidence from a corpora of biomedical literature), RobotReviewer [Marshall *et al.*, 2016; Marshall *et al.*, 2017] (which summarizes the key information of a clinical trial, including the interventions, trial participants and risk of bias), and ExaCT [Kiritchenko *et al.*, 2010] (which extracts information containing PICO elements based on a SVM). Also, Lehman et al. [Lehman *et al.*, 2019] proposed an approach to infer if a study provides evidence with respect to a given intervention, comparison intervention and outcome. However, none of these systems is able to extract a full argument graph (where evidence and claims are the nodes, and attacks and supports are the labelled edges) from a clinical text. Concerning the identification of PICO elements in text, different approaches are proposed in the literature [Dhrangadhariya *et al.*, 2021; Jin and Szolovits, 2018; Trenta *et al.*, 2015] to identify them in text, but none of these approaches tackles the issue of analysing the effects of an intervention on the outcomes of a clinical trial study, as in ACTA 2.0.

2 ACTA 2.0 Main Functionalities

ACTA 2.0 provides the following functionalities, visualized in Figure 1:

Search on Pubmed. PubMed[2] is a free search engine accessing primarily the MEDLINE database[3] of references and

*Contact Author
[1]http://ns.inria.fr/acta/
[2]https://pubmed.ncbi.nlm.nih.gov/
[3]https://www.nlm.nih.gov/medline/medlineoverview.html

abstracts on life sciences and biomedical topics. Given the importance of this search engine in the health-care domain, ACTA 2.0 maintains the possibility to search for a (set of) abstract(s) directly on the PubMed catalogue, through their API[4]. As in the previous version of ACTA, this API is integrated as a search bar to enter queries in the common PubMed format, similarly to the original PubMed web interface. After the query is executed, when the results are shown, the user can then select one or more abstracts to proceed with the analyses offered by ACTA 2.0. Alternatively, the system accepts raw text as input to be processed via *Analyse Custom Text*.

Enhanced Argumentative Analysis. Once the text is uploaded or an abstract is selected from the list of search results, the user can proceed with the argumentative and outcome analyses by pressing the *Analyse* button. After a short processing time, the result is visualized in the user interface in form of an argumentative graph. There, the nodes are the premises and the claims automatically detected in the abstract, and the labelled edges correspond to the relations among them. In contrast to the previous version, ACTA 2.0 integrates a completely overhauled relation classification module, implementing the methods described by Mayer *et al.* [Mayer *et al.*, 2021]. Besides underlying technical changes regarding architecture, loss function and problem formulation, the most notable difference for the user is the updated linking of the arguments in the graph, which are now not only identified, but also labeled, indicating their argumentative function as either *attack* or *support*. For readability purposes, the text of the argumentative components is not shown by default in the argumentation graph. However, the user can unveil it by interacting with the graph, i.e., hovering over the respective argument component. Additionally, argument components are highlighted with different colors (evidence in blue, claims in orange) in the abstract, which is always fully shown on the right side of the window.

PICO Element Detection. The detected PICO elements can be visualized in a similar fashion through the *PICO Elements* button. Again, each PICO category is highlighted in a different color. For the PICO detection, we rely on the same module employed in the first version of ACTA.

Effects on Outcomes. As one of the major upgrades, ACTA 2.0 implements a new module to analyse the reported effects an intervention has on the outcomes (**O** of PIC**O**) in the clinical trial abstract. Such as if an intervention increased or decreased the measured outcome, as proposed in [Mayer *et al.*, 2021]. The underlying motivation for this is twofold: first, to enrich the arguments with valuable medical information and thus increase versatility of the application; and second, to provide structured and machine-processable data, which can serve as input to a computational model of argument system [Atkinson *et al.*, 2017], for instance. In the web interface of the tool, these effects can also be visualized by pressing the *Effects on Outcome* button. As a consequence, the outcomes are highlighted in the displayed abstract to the right with different colors according to their predicted effect, i.e., *Increased*, *Decreased*, *Improved*, *NoOccurrences* or *NoDifferences*.

ACTA 2.0 Public API. Another major upgrade is the conversion from a static monolithic pipeline to a modular and extendable system to foster versatility and re-usability. In particular, each of the processing steps, i.e., argument component detection, relation classification, PICO and effect prediction, are now independent executable units, which can be called separately via our publicly available REST API[5]. Researchers, developers and clinicians can now not only try them all individually, but have also the possibility to replace or add custom modules to the workflow, or build parts into their own projects. The required input and output formats for each module are defined in the documentation of the API.

Data Format Description. Each module takes as input a JSON file, where for the argument components, the PICO elements and Outcome Detection modules, the field "text" must be filled in with the medical text to be analyzed. For the relation classification module, the input JSON file must have the field "candidates" filled with the list of all of the argumentative components text and type (claim or premise) for which the user wants to predict the relation (support or attack). For the effect prediction module, both the original text and the selected outcomes have to be provided in the "text" and "outcomes" fields respectively. For every module, a JSON file is produced as output with the corresponding results, either being the detected component spans or the predicted labels. All the results, including the argumentative analysis together with PICO elements and effects on outcomes, can be downloaded as a JSON file for each of the processed abstracts.

3 Experimental Setting and Results

Argumentative Analysis and PICO Extraction. For the argumentative analysis and the PICO element detection, we replicate the same setting of Mayer *et al.* [Mayer *et al.*, 2019], evaluating ACTA 2.0 on the same sample of the AbstRCT dataset[6] (i.e., 500 abstracts of randomized controlled trials on neoplasm treatment annotated with claim, premise and their relations). We use the BIO-tagging scheme for the sequence tagging problem of the argumentative component detection, based on a pre-trained bidirectional transformer language model [Mayer *et al.*, 2021]. We thus keep the BERT base [Devlin *et al.*, 2019] model for the token-level representation of contextualized sentences and entirely fine-tune it during three epochs with an Adam optimizer and a learning rate of 2e-5. The sentence representation is passed into a Recurrent Neural Network, here a Gated Recurrent Unit (GRU [Cho *et al.*, 2014]) and then to a Conditional Random Field (CRF [Lafferty *et al.*, 2001]). The model archives a f1-score of 85.2 on the neoplasm test set. PICO element detection employs the same methods to train the model on the EBM-NLP dataset [Nye *et al.*, 2018] with coarse labels. The dataset splits are the same than in [Mayer *et al.*, 2021] without sentences containing less than 10 WordPiece [Wu *et al.*, 2016]. The obtained f1-score on the test set is 73.4.

[4]https://pubmed.ncbi.nlm.nih.gov/advanced/

[5]https://ns.inria.fr/acta/doc/

[6]https://gitlab.com/tomaye/abstrct/

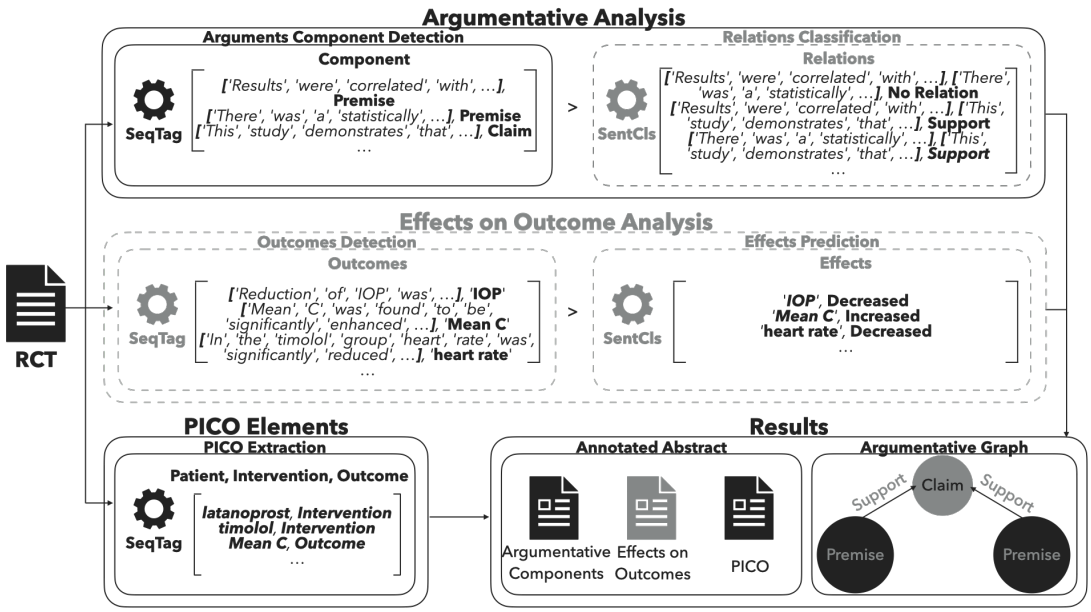

Figure 1: ACTA 2.0 pipeline (the newly introduced modules are in grey).

Class	#outcomes	%
Improved	831	25
Increased	765	23
Decreased	782	23
NoDifference	897	27
NoOccurrence	76	2
Total	3351	100

Table 1: Statistics of the Outcome dataset, showing the numbers of *Improved*, *Increased*, *Decreased*, *NoDifference* and *NoOccurrence* classes independent of the disease-based subsets.

Labeled Relation Classification. For relation classification, we rely on a bi-directional transformer, but we change the representation of the sequence classification problem jointly modelling the relations by classifying all the argumentative component combinations. This new representation is passed to a linear layer with a softmax which classifies it into the three target classes (*Support*, *Attack* and *NoRelation*). The SciBERT [Beltagy *et al.*, 2019] uncased base model with pre-trained weights is used for the sentence representations, fine-tuned with a learning rate of 2e-5, batch size of 8, maximum sentence length of 256 sub-words tokens per input example during 3 epoch. The weight factor for each of the 3 classes in the weight cross entropy loss is the normalized number of training samples of this class. This settings archive respectively a macro f1-score of 0.68, 0.70 and 0.70 for the Neoplasm, Glaucoma and Mixed test sets [Mayer *et al.*, 2021].

Effects on Outcomes. To evaluate the Effects on Outcome task, we focus on the sentences of the AbstRCT dataset containing outcomes (i.e., 3351 sentences annotated with five classes, as reported in Table 1). Such sentences are shuffled and split respecting the class distributions to build the 80% and 20% train and test sets, respectively. The task is addressed as a two-step pipeline: *(i)* outcome detection, and *(ii)* effect prediction [Mayer *et al.*, 2021]. The first sub-task is casted as a sequence tagging problem using the 3 class BIO-tagging scheme to detect the outcome boundaries. We use a pre-trained bidirectional transformer language model trained on scientific data, i.e., SciBERT uncased, fine-tuning it for three epochs with the same hyper-parameters we used in the argument component detection module. We then pass the sentence representations across a Gated Recurrent Unit and a Conditional Random Field. We then address the effect prediction subtask as sequence classification, where each outcome together with the component it occurred in is provided as input into the effect classifier. The same pre-trained transformer model types as for relation classification (based on SciBERT combined to a bidirectional GRU and a final CRF) are used to predict one among the fives effect classes in Table 1. The outcome detection and effect classification tasks together reach a macro f1-score of 0.80.

4 Concluding Remarks

We presented a new version of the ACTA system for the argumentative analysis of clinical trial. ACTA 2.0 allows for the automatic extraction and interactive visualization of the arguments contained in the abstracts of clinical trial articles on PubMed, and the automatic analysis of the effects on the outcomes concerning the extracted PICO elements. A REST API is provided to allow for the embedding of the single services in other systems. Further improvements include the automatic prediction of the link between the detected outcome and the related intervention(s), and the analysis of a set of abstracts together in such a way to identify the relations among different arguments belonging to different abstracts.

Acknowledgements

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. This work was supported by the CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4-0002.

References

[Atkinson *et al.*, 2017] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Mag.*, 38(3):25–36, 2017.

[Beltagy *et al.*, 2019] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3615–3620, 2019.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, 2019.

[Dhrangadhariya *et al.*, 2021] Anjani Dhrangadhariya, Gustavo Aguilar, Thamar Solorio, Roger Hilfiker, and Henning Müller. End-to-end fine-grained neural entity recognition of patients, interventions, outcomes. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 65–77, 2021.

[Jin and Szolovits, 2018] Di Jin and Peter Szolovits. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of BioNLP 2018 workshop*, pages 67–75, 2018.

[Kiritchenko *et al.*, 2010] Svetlana Kiritchenko, Berry de Bruijn, Simona Carini, Joel Martin, and Ida Sim. ExaCT: Automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10:56, 2010.

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*, pages 282–289, 2001.

[Lehman *et al.*, 2019] Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of NAACL-HLT 2019*, pages 3705–3717, 2019.

[Marshall *et al.*, 2016] Iain J Marshall, Joël Kuiper, and Byron C Wallace. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201, 2016.

[Marshall *et al.*, 2017] Iain Marshall, Joël Kuiper, Edward Banner, and Byron C. Wallace. Automating biomedical evidence synthesis: RobotReviewer. In *Proceedings of ACL 2017, System Demonstrations*, pages 7–12, 2017.

[Mayer *et al.*, 2019] Tobias Mayer, Elena Cabrio, and Serena Villata. ACTA a tool for argumentative clinical trial analysis. In *Proceedings of IJCAI-19*, pages 6551–6553, 7 2019.

[Mayer *et al.*, 2021] Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artificial Intelligence in Medicine*, page 102098, 2021.

[Nye *et al.*, 2018] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of ACL 2018*, pages 197–207, 2018.

[Sackett and Rosenberg, 1995] David L. Sackett and William M. C. Rosenberg. On the need for evidence-based medicine. *Journal of Public Health*, 17(3):330–334, 1995.

[Trenta *et al.*, 2015] Antonio Trenta, Anthony Hunter, and Sebastian Riedel. Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints. *CoRR*, abs/1509.05209, 2015.

[Wang *et al.*, 2020] Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, and Jiawei Han. EVIDENCEMINER: Textual evidence discovery for life sciences. In *Proceedings of ACL 2022: System Demonstrations*, pages 56–62, 2020.

[Wu *et al.*, 2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.