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

AMICA: An Argumentative Search Engine for COVID-19 Literature

Marco Lippi[1,*], **Francesco Antici**[2], **Gianfranco Brambilla**[3], **Evaristo Cisbani**[3], **Andrea Galassi**[2], **Daniele Giansanti**[3], **Fabio Magurano**[3], **Antonella Rosi**[3], **Federico Ruggeri**[2] and **Paolo Torroni**[2]

[1]DISMI – University of Modena and Reggio Emilia, Italy
[2]DISI – University of Bologna, Italy
[3]National Institute of Health, Italy

Abstract

AMICA is an argument mining-based search engine, specifically designed for the analysis of scientific literature related to COVID-19. AMICA retrieves scientific papers based on matching keywords and ranks the results based on the papers' argumentative content. An experimental evaluation conducted on a case study in collaboration with the Italian National Institute of Health shows that the AMICA ranking agrees with expert opinion, as well as, importantly, with the impartial quality criteria indicated by Cochrane Systematic Reviews.

1 Introduction

One effect of the COVID-19 pandemics is undoubtedly a huge amount of novel scientific literature. A significant part of it is made of *preprints*, that is, papers made publicly available *before* peer-reviewing. As a matter of fact, preprints have become crucial in explosive situations, such as the one in question, that require a continued rapid sharing of information within the scientific community [Bedford *et al.*, 2020]. As a result, in an effort to capture relevant and anticipatory topics, an overwhelming amount of literature has to be carefully studied and analyzed by domain experts, who also have to face the challenge of filtering out low-quality papers in a timely and accurate fashion. It quickly became clear that experts alone cannot deal with the management of huge collections of data: automatic tools need to be constructed, that can help detecting and extracting the most relevant pieces of information [Brainard, 2020]. Contributors from diverse areas of artificial intelligence have thus proposed initiatives to encourage the development of enabling technologies and resources such as COVID-19 scientific paper datasets,[1] and various machine learning and natural language processing-based platforms and tools [Verspoor *et al.*, 2021; Menin *et al.*, 2021].[2]

In this paper, we describe a prototype intelligent system designed to address the challenge of automatic analysis of scientific literature related to COVID-19 through argument mining (AM) [Lippi and Torroni, 2016a]. AM is a research field at the intersection of natural language processing, computational linguistics, argumentation, logic, and philosophy. AM aims to build systems that can automatically extract arguments from natural language documents. Most of the existing AM tools are tailored to specific domains and genres, but a few general-purpose systems and tools are available. One of them is MARGOT [Lippi and Torroni, 2016b], which was successfully applied to different types of documents, like clinical trials [Mayer *et al.*, 2018], the grey literature in software engineering research [Williams, 2019], and Amazon reviews [Passon *et al.*, 2018]. The whole field of AM is rapidly evolving, and there is a general effort in developing novel instruments for the end-user [Lytos *et al.*, 2019]. Recently, some preliminary studies have also reported the application of AM tools to the analysis of the scientific literature related to COVID-19 [Menin *et al.*, 2022].

The key idea of AMICA is that of a search engine that can automatically identify not just scientific articles, but *arguments* in scientific articles, that are relevant to a user query. Given a statement or a sequence of keywords (the *query*), AMICA will assign a higher rank to papers that are not only related with the query, but also have a richer argumentative content. We shall remark that this idea is not specific to the COVID-19 scientific literature. On the contrary, the AMICA tool can be applied to other domains of medical and scientific literature. In a broader sense, AMICA is an innovative system in that it uses AM tools as a key enabler for the detection of arguments within large, unstructured document collections. Arguments include *claims*, i.e., controversial or debatable statements regarding a certain topic of interest (e.g., "In hospitalized adult patients with severe COVID-19, no benefit was observed with lopinavir–ritonavir treatment beyond standard care"). Claims can be further supported by premises, sometimes also called *evidence* ("A total of 199 patients with laboratory-confirmed SARS-CoV-2 infection underwent randomization; 99 were assigned to the lopinavir–ritonavir group, and 100 to the standard-care group.").

Our underlying hypothesis is that claims and evidence of the kind illustrated above can be effectively and automatically extracted from a large amount of medical research articles and reports, and that a presentation of a selected number of such arguments may help medical researchers better than argument-agnostic text mining tools that do not distinguish

*Contact author: marco.lippi@unimore.it
[1]https://www.semanticscholar.org/cord19
[2]https://www.kaggle.com/covid-19-contributions

between claims/evidence and other types of information.

2 System Implementation

AMICA is deployed as a web-based search engine. It receives as input a user query, such as a set of keywords, or a statement regarding a certain topic, and it returns a set of documents, ranked by their relevance with respect to the keyword, as well as by their argumentative content. In what follows, we will briefly illustrate the AM tools used by AMICA, the data collection procedure for building the scientific papers database, and the scoring function exploited by AMICA in the document ranking phase.

2.1 MARGOT

AMICA relies on a freely-available system that identifies claims/evidence from a given input text, called MARGOT [Lippi and Torroni, 2016b].[3] In particular, MARGOT assigns a claim score CS and an evidence score ES to each of its sentences, representing its confidence that the sentence contains a claim or evidence, respectively. MARGOT is based on tree kernels [Moschitti, 2006] and was trained on the AM corpus developed by IBM in the context of the Debater project, which consists of Wikipedia pages. An extensive study demonstrated that MARGOT can generalize across different genres and styles [Lippi and Torroni, 2016b].

2.2 Dataset Collection

The AMICA search engine relies on a dataset of documents to be processed offline using MARGOT. The dataset is updated with new preprints on a regular basis. To that end, AMICA implements a pipeline where a Python script fetches research papers and stores their relevant information (metadata) in a MongoDB database, thus keeping the query response time within reasonable usability bounds. Papers are retrieved from open access archives that freely provide dedicated APIs, such as arXiv, bioRxiv, medRxiv, and Research Square. In order to be interpreted and displayed by the AMICA web platform, each paper needs to undergo several pre-processing steps:

1. *Parsing*: we defined a fixed global structure to store papers leading to the development of a series of rules to parse each document, according to its source.
2. *Filtering*: papers already present in the database are filtered out in order to avoid duplicates, via a matching criterion on the database queries.
3. *Text Extraction*: after gaining access to the pdf version of the new paper, we exploited an open source tool called Grobid[4] to extract plain text from it.
4. *MARGOT annotations*: having obtained the plain text of the document, MARGOT can be run on it, so that claim and evidence scores can be stored as metadata for efficient retrieval purposes.

Following this approach, we collected an initial corpus of over 15,000 papers, that have been already processed by MARGOT for argument detection, and that can be used as the search database by AMICA. The corpus is continuously updated on a weekly basis. The full implementation of the described pipeline is publicly available at the following repository: https://github.com/francescoantici/amica.

[3] http://margot.disi.unibo.it

[4] https://github.com/kermitt2/grobid

Figure 1: Results page provided by the AMICA server.

2.3 Scoring Function

Another key ingredient of AMICA is the scoring function used to rank papers. For the purposes of AMICA, given a query q and a set of n documents $\mathcal{D} = \{d_1, \ldots, d_n\}$, a scoring function should take into account not only the similarity between documents and query, but also the argumentative score $AS(d_j)$ of each document d_j. In other words, the scoring function $s_a(q, d_j)$ should assign a high value to documents that are both relevant for the query q and rich in terms of argumentation. For each sentence, we chose to pick the maximum between the claim score CS and the evidence score ES, since in order to be considered argumentative, a sentence has to contain either a claim or evidence, but doesn't necessarily have to contain both. Then, we aggregate over an entire document by averaging over the set of sentences N_{d_j}:

$$AS(d_j) = \frac{1}{N_{d_j}} \sum_{k=1}^{N_{d_j}} \max\{CS(sent_k), ES(sent_k)\} \quad (1)$$

Conversely, classic information retrieval systems use a scoring function $s(q, d_j)$ that computes a similarity between q and d_j. A typical similarity score is the cosine similarity between the bag-of-words vectors or dense embeddings representing q and d_j. AMICA combines the argumentative score AS, provided by MARGOT independently of the query, with the classic similarity score $s(q, d_j)$ based on bag-of-words, by multiplying the two factors. The AMICA score s_a is thus:

$$s_a(q, d_j) = s(q, d_j) \cdot AS(d_j) \quad (2)$$

More sophisticated scores could be conceived. However, our experimental evaluation indicates that this solution is not only simple and efficient, but also effective.

2.4 Web Interface

The AMICA system prototype is at http://amica.unimore.it. The home page features an input form for the query, and a button to run the analysis. Figure 1 shows a screenshot of the results web page. For each paper, AMICA shows a link to the pdf, a link to the analysis performed by MARGOT, a brief snippet, and the two scores $s(q, d_j)$ and $AS(d_j)$ employed in the ranking function. A brief video describing the system is available at http://amica.unimore.it/AMICA_video.mov.

3 Experimental Evaluation

We evaluated AMICA by running two experiments. First, we compared the scores produced by the AMICA system with the impartial quality criteria provided by a Cochrane Systematic Review,[5] which are standardized, high-quality reviews that systematically analyze a given topic, providing a list of eligibility criteria and conditions to evaluate the methodology and the empirical evidence provided by scientific papers. Second, we compared the AMICA scores with relevance scores provided by a pool of domain experts, to assess the correlation between the scores. In this study, the experts were four researchers of the Italian National Institute of Health.

As a case study, we considered a Cochrane Systematic Review on rapid, point-of-care and molecular-based tests for the diagnosis of COVID-19 [Dinnes *et al.*, 2021]. We collected 40 papers from those surveyed in the review, 20 of which were included and 20 of which were excluded according to the Cochrane eligibility criteria. We extracted the argument components of these 40 papers with MARGOT, and we computed a set of statistics describing the argumentative content of each paper. In particular, we computed the argument ratio AR as the percentage of sentences containing at least one argument component, and the argumentative score AS.

Figure 2 shows a scatter plot where each point represents a document, reporting the AR and AS statistics on the two axes. Papers included in the Cochrane Systematic Review are displayed in orange, whereas those excluded by the review are coloured in blue. The diagram shows an evident correlation between the amount of argumentative content and the inclusion in the Cochrane: the largest part of papers included in the Cochrane review, in fact, are also the most argumentative papers for MARGOT (larger AR and AS, upper-right corner). Conversely, the lower-left corner, with the least argumentative papers, contains almost entirely papers that are excluded by the review.

In order to validate the ranking induced by MARGOT with that proposed by a pool of experts, we asked four domain experts to give a score from 1 to 5 to the 40 papers.[6] We considered the average of the scores collected for each paper, and we computed the Spearman's rank correlation coefficient ρ against the statistics computed by MARGOT. We obtained $\rho = 0.463$ when considering the AS score for MARGOT, and $\rho = 0.526$ when taking into account the AR score, which corresponds to a moderate-to-strong correlation [Akoglu, 2018].

[5]https://www.cochranelibrary.com/cdsr/about-cdsr

[6]The experts agreed on a set of guidelines to align their scores.

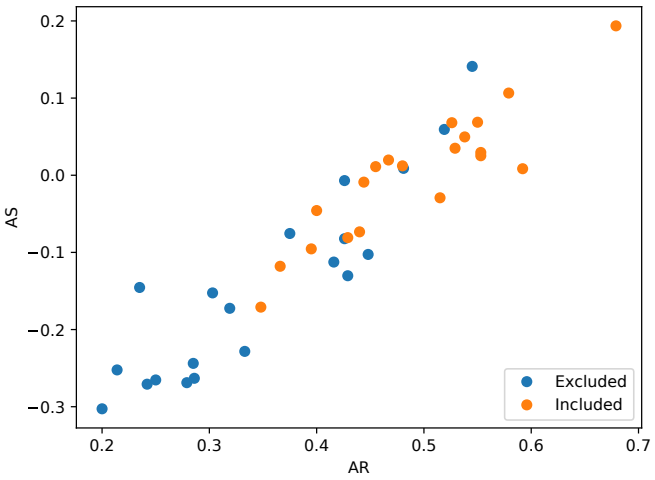

Figure 2: Scatter plot of the argumentative content of 40 papers analyzed by a Cochrane Systematic Review on rapid, point-of-care and molecular-based diagnostic tests for COVID-19. AR and AS are the argument ratio and the argumentative score, respectively, associated by MARGOT to each paper. Blue dots correspond to papers that were excluded from the review, whereas orange ones correspond to papers that were included.

4 Conclusions

In the last two years, the stream of scientific papers related to COVID-19 has grown dramatically. An urgent need ensued, for automated systems able to scan this overwhelming amount of information and provide quality indicators to medical experts, healthcare professionals, scientists and policy-maker. This phenomenon does not apply only to COVID-19, but also to many other topics in the domains of science and healthcare.

To help sifting through this huge, growing collection of scientific documents, we developed the AMICA system: a search engine based on argument mining techniques, which scores papers based on their relevance to a query as well as on their argumentative content. An experimental evaluation conducted in collaboration with the Italian National Institute of Health confirms the relevance of the ranking produced by AMICA with respect to the score provided by domain experts. The system is freely available as a web server.

In the future, we plan to extend our approach to other domains in biomedical literature, and to compare the results provided by AMICA also to the ranking produced by other search engines.

Acknowledgments

The research conducted in this paper was supported by the Italian Ministry for Education and Research, under the FISR-COVID 2020 project "AMICA".

References

[Akoglu, 2018] Haldun Akoglu. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93, 2018.

[Bedford *et al.*, 2020] Juliet Bedford, Delia Enria, Johan Giesecke, David L Heymann, Chikwe Ihekweazu, Gary Kobinger, H Clifford Lane, Ziad Memish, Myoung-don Oh, Anne Schuchat, et al. Covid-19: towards controlling of a pandemic. *The lancet*, 395(10229):1015–1018, 2020.

[Brainard, 2020] Jeffrey Brainard. Scientists are drowning in covid-19 papers. can new tools keep them afloat. *Science*, 13(10.1126), 2020.

[Dinnes *et al.*, 2021] Jacqueline Dinnes, Jonathan J Deeks, Sarah Berhane, Melissa Taylor, Ada Adriano, Clare Davenport, Sabine Dittrich, Devy Emperador, Yemisi Takwoingi, Jane Cunningham, et al. Rapid, point-of-care antigen and molecular-based tests for diagnosis of sars-cov-2 infection. *Cochrane Database of Systematic Reviews*, 3(3), 2021.

[Lippi and Torroni, 2016a] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25, 2016.

[Lippi and Torroni, 2016b] Marco Lippi and Paolo Torroni. Margot: A web server for argumentation mining. *Expert Systems with Applications*, 65:292–303, 2016.

[Lytos *et al.*, 2019] Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management*, 56(6):102055, 2019.

[Mayer *et al.*, 2018] Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. Argument mining on clinical trials. In *COMMA*, pages 137–148, 2018.

[Menin *et al.*, 2021] Aline Menin, Franck Michel, Fabien Gandon, Raphaël Gazzotti, Elena Cabrio, Olivier Corby, Alain Giboin, Santiago Marro, Tobias Mayer, Serena Villata, et al. Covid-on-the-web: Exploring the covid-19 scientific literature through visualization of linked data from entity and argument mining. *Quantitative Science Studies*, pages 1–42, 2021.

[Menin *et al.*, 2022] Aline Menin, Franck Michel, Fabien Gandon, Raphaël Gazzotti, Elena Cabrio, Olivier Corby, Alain Giboin, Santiago Marro, Tobias Mayer, Serena Villata, and Marco Winckler. Covid-on-the-Web: Exploring the COVID-19 scientific literature through visualization of linked data from entity and argument mining. *Quantitative Science Studies*, 2(4):1301–1323, 02 2022.

[Moschitti, 2006] Alessandro Moschitti. Making tree kernels practical for natural language learning. In *11th conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[Passon *et al.*, 2018] Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso. Predicting the usefulness of amazon reviews using off-the-shelf argumentation mining. In Noam Slonim and Ranit Aharonov, editors, *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 35–39. Association for Computational Linguistics, 2018.

[Verspoor *et al.*, 2021] Karin Verspoor, Simon Šuster, Yulia Otmakhova, Shevon Mendis, Zenan Zhai, Biaoyan Fang, Jey Han Lau, Timothy Baldwin, Antonio Jimeno Yepes, and David Martinez. Brief description of covid-see: The scientific evidence explorer for covid-19 related research. In *European Conference on Information Retrieval*, pages 559–564. Springer, 2021.

[Williams, 2019] Ashley Williams. *Finding high-quality grey literature for use as evidence in software engineering research.* PhD thesis, University of Canterbury, 2019.