Overview of Touché 2021: Argument Retrieval
Extended Abstract

Alexander Bondarenko[1](✉), Lukas Gienapp[2], Maik Fröbe[1], Meriem Beloucif[3], Yamen Ajjour[1], Alexander Panchenko[4], Chris Biemann[3], Benno Stein[5], Henning Wachsmuth[6], Martin Potthast[2], and Matthias Hagen[1]

[1] Martin-Luther-Universität Halle-Wittenberg, Halle, Germany
touche@webis.de
[2] Leipzig University, Leipzig, Germany
[3] Universität Hamburg, Hamburg, Germany
[4] Skolkovo Institute of Science and Technology, Moscow, Russia
[5] Bauhaus-Universität Weimar, Weimar, Germany
[6] Paderborn University, Paderborn, Germany
https://touche.webis.de

Abstract. Technologies for argument mining and argumentation analysis are maturing rapidly, so that, as a result, the retrieval of arguments in search scenarios becomes a feasible objective. For the second time, we organize the Touché lab on argument retrieval with two shared tasks: (1) argument retrieval for controversial questions, where arguments are to be retrieved from a focused debate portal-based collection and, (2) argument retrieval for comparative questions, where argumentative documents are to be retrieved from a generic web crawl. In this paper, we briefly summarize the results of Touché 2020, the first edition of the lab, and describe the planned setup for the second edition at CLEF 2021.

1 Introduction

Making informed decisions and forming personal opinions are everyday tasks, requiring one to choose between two or more options, or sides. This may be based on prior knowledge and experience, but more often requires the collection of new information first. The web is rife with documents comprising arguments and opinions on many controversial topics, as well as on products, services, etc. However, hardly any support is provided for individuals who specifically search for argumentative texts in order to support their decision making or opinion formation. Especially for controversial topics, search results are often riddled with populism, conspiracy theories, and one-sidedness, all of which arguably do not lead to the kind of insights that help individuals form well-justified informed opinions. But even straightforward tasks, such as comparing among two specific options for a product, are sometimes challenging to be solved with a web search

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, pp. 574–582, 2021.
https://doi.org/10.1007/978-3-030-72240-1_67

engine, given that advertisers optimize their sales pages and compete with others for the top-most search result slots, displacing reasoned comparisons.

To foster research on argument retrieval in the scenarios of (1) opinion formation on controversial topics, and (2) personal “everyday” decision making, we organize the second edition of the Touché lab on Argument Retrieval at CLEF 2021.[1] Participants of the lab are asked to develop a technology that helps to retrieve “strong” arguments for decisions at the societal level (e.g., “Is climate change real and what to do?”) and at the personal level (e.g., “Should I buy real estate or rent, and why?”). The corresponding two shared tasks are:

1. *Argument Retrieval for Controversial Questions.* Argument retrieval from a focused document collection (crawled from debate portals) to support opinion formation on controversial topics.
2. *Argument Retrieval for Comparative Questions.* Argument retrieval from a generic web crawl to support decision making in “everyday” choice situations.

Our goal is to establish an understanding of how to evaluate argument retrieval and what retrieval models or processing methods are effective. For instance, an important component of argument retrieval probably is the assessment of argument quality (i.e., whether a given argument is a “strong” one). Good argument retrieval approaches will not only allow for a better handling of argumentative information needs in search engines, but they may also become, in the long run, an enabling technology for automatic open-domain agents that convincingly discuss and interact with human users.

2 Task Definition

The Touché lab adopts the standard TREC-style setup and evaluation methodology, where document collections and a set of search topics are provided to the participants. Every topic is comprised of a search query, a detailed description of the search scenario, and hints on document relevance for assessors.

The second edition of the lab repeats the two shared tasks from the first edition with specific twists: (1) Topics and judgments from the first year are made available to the participants for training their models,[2] (2) new topics are composed and used for the evaluation of the submitted approaches, and (3) in addition to relevance, several argument quality dimensions are evaluated (cf. Sect. 2.4).

To allow for a high diversity of approaches, participating teams are allowed to submit up to five runs (differently ranked result lists) that then form part of the judgment pool passed to expert assessors. We encourage the participating teams to submit the software implementing their approaches within the evaluation platform TIRA [10] in order to maximize the reproducibility.

[1] ‘Touché’ is commonly “used to acknowledge a hit in fencing or the success or appropriateness of an argument, an accusation, or a witty point.” [https://merriam-webster.com/dictionary/touche].

[2] Available for download on the lab website: https://touche.webis.de.

2.1 Task 1: Argument Retrieval for Controversial Questions

The first shared task focuses on the scenario of supporting individuals who search directly for arguments on controversial topics of general societal interest (e.g., immigration, climate change, or the use of plastic bottles). The retrieved arguments relevant to such topics should be useful in debating conversations, or be helpful in forming an opinion on the topic. Multiple online portals are centered around argumentative topics (e.g., debate portals), yet general web search engines do not offer an effective way to retrieve "strong" arguments from these platforms (cf. Sect. 2.4 for a more detailed description of argument strength). However, there are some prototypes of argument search engines, such as args.me [17], ArgumenText [14], and TARGER [5], that implement different paradigms to solve the task of argument retrieval. While the args.me approach first identifies a focused collection of arguments crawled from online debate portals and then indexes only these arguments, ArgumenText and TARGER follow a more "traditional" web-based retrieval approach and mine arguments from a query's result documents in a post-processing step.

In Task 1, to ensure a low entry barrier for participants, we use the existing args.me argument corpus [1]. This corpus is a focused crawl obtained from four online debate portals (idebate.org, debatepedia.org, debatewise.org, and debate.org) and thus mostly contains short and to-the-point arguments exchanged during an online debate. This way, participants of Task 1 do not necessarily need to manage a fully-fledged argument mining pipeline for participation. The corpus is available for download and can also be queried directly using the API of args.me [17].[3]

2.2 Task 2: Argument Retrieval for Comparative Questions

The second shared task aims to support users in personal decisions when choosing between different options. In particular, the task is to find relevant documents containing "strong" arguments for questions like "Is X better than Y for Z?". In their current form, web search engines do not provide much support for such comparative questions; they even sometimes retrieve one-sided answers from community question answering platforms. A state-of-the-art system to deal with comparative information needs is the comparative argumentation machine CAM [13], which takes two objects to be compared as well as a set of aspects of comparison as input, retrieves comparative sentences in favor of one or the other option from a 2016 Common Crawl version using BM25 as the retrieval model, and clusters the sentences to present a summary table. However, CAM cannot process queries represented as questions, it processes relevant information only on sentence level, and it does not account for argumentative aspects of answers. Improving retrieval models for systems like CAM is the objective of Task 2.

[3] https://www.args.me/api-en.html.

The participants of Task 2 are asked to retrieve documents from the general-purpose web crawl ClueWeb12[4] to help individuals come to an answer for some comparative question. Ideally, relevant documents should comprise "strong" arguments for or against one or the other option underlying the comparative question. For participants who do not want to index the ClueWeb12 at their site, a retrieval functionality is made accessible via the API of the (argumentation-agnostic) reference search engine ChatNoir [2].[5] Furthermore, the APIs of argument tagging tools like Targer [5][6] may be used to identify argumentative units (i.e., claims and premises) in free text input.

2.3 Search Topics and Training Data

For each shared task, we provide 100 search topics (50 from the first lab edition for training and 50 new topics for evaluation), ensuring that respective information can be found in the focused crawl of debate portals and in the ClueWeb12, respectively. Every topic consists of (1) a *title* representing a question on some controversial topic or some choice problem, (2) a *description* providing a detailed definition of the respective scenario, and (3) a *narrative* that is part of the "guideline" used for relevance and argument quality labeling by expert assessors. The topics (previous and new) and the relevance judgments from the first lab edition (for 5,262 unique arguments in Task 1 and for 1,783 unique documents in Task 2) are available to the participants—to, for instance, allow training or fine-tuning of (neural) retrieval models. The participants' submitted ranked document lists (runs) are also available to analyze the submitted rankings. Given the relatively small training data, the participants may of course also exploit document relevance judgments collected at other shared tasks (e.g., various TREC tracks[7]) that, for instance, can be found in the Anserini GitHub repository [18].[8] Additionally, for argument quality assessment, corpora such as the ones published by Gienapp et al. [6], Gretz et al. [7], Toledo et al. [15], or Wachsmuth et al. [16] may be used.

2.4 Evaluation

The evaluation is based on the pooled top results of the participants' submitted runs. For these, human assessors label argumentative text passages or documents manually, both for their general topical relevance, and for argument quality dimensions found to be important for the evaluation of arguments [16].

For Task 1, we assess three such quality dimensions: whether an argumentative text is logically cogent, whether it is rhetorically well-written, and whether it contributes to the users' stance-building process (i.e., "dialectical quality",

[4] https://lemurproject.org/clueweb12/.
[5] https://www.chatnoir.eu/doc/api/.
[6] https://demo.webis.de/targer-api/apidocs/.
[7] https://trec.nist.gov/tracks.html.
[8] https://github.com/castorini/anserini/.

similar to the concept of "utility") [16]. For Task 2, in addition to a document's relevance, human assessors judge whether sufficient argumentative support is provided as defined by Braunstain et al. [4], and they evaluate the credibility of web documents as defined by Rafalak et al. [11]. Thus, a "strong" argument is defined as one that fulfills certain criteria of argument quality such as logical cogency, rhetorical quality, contribution to stance-building, level of support, and credibility. The studies carried out by Potthast et al. [9] and Gienapp et al. [6] suggest that argument quality assessment is feasible also via crowdsourcing, yet, challenging for untrained annotators. For this reason, we specifically pay attention to developing annotation instructions and quality control instruments (pilot judgments, assessing inter-annotator agreement, and recruiting assessors externally and internally).

The effectiveness of the participants' submitted approaches is measured in traditional ranking-based ways with respect to relevance and the qualitative aspects of arguments (e.g., nDCG [8] using the graded relevance or quality judgments).

3 Touché at CLEF 2020: Results and Findings

In the first edition of the Touché lab, 28 teams registered, from which 17 actively participated in the shared tasks by submitting approaches/results [3]. The majority of the participating teams used the TIRA platform [10] to submit software that then produced runs after being invoked at our site. The run output files follow the standard TREC-style format. The teams were allowed to submit several runs, but asked to give evaluation priorities in case more than one run was submitted. This resulted in 41 valid submitted runs from the 17 teams. From every team, at least the top five runs of highest priority were pooled for further evaluation. Additionally, we included rankings produced by two baseline systems in the evaluation: the Lucene implementation of query likelihood with Dirichlet-smoothed language models (DirichletLM [19]) for Task 1, and the BM25F-based [12] search engine ChatNoir [2] for Task 2. We briefly summarize the main results and findings here; the lab overview contains more specific information [3].

Task 1: Argument Retrieval for Controversial Questions. The submissions to Task 1 (13 teams submitted 31 runs, one additional baseline) mainly followed a general strategy consisting of three components: (1) a retrieval model, (2) an augmentation (either query expansion or an extension of an initially retrieved result set), and (3) a (re-)ranking approach based on some document features that boosted or modified the initial retrieval scores, or that were used directly to rank the initial results.

Most of the participating teams chose one of four retrieval models. In the evaluation, DirichletLM and DPH were much more effective than BM25 and TF-IDF. Half of the submitted approaches opted to integrate query or result augmentation by applying various strategies. Queries were expanded by synonyms or they were augmented with newly generated queries using large pre-trained language models. The retrieval results were sometimes post-processed using argument clustering (e.g., topic models or semantic clustering) and in a final re-ranking step, the majority of the approaches either exploited some notion of argument quality or utilized sentiment analysis. Other re-ranking features include premise prediction scores, text readability, presence of named entities, and credibility scores of argument authors in the corpus.

In the first lab edition, we evaluated only the relevance of the retrieved arguments (not their quality) using the nDCG [8] implementation provided by the *trec_eval* library[9] with an evaluation depth of five. Following annotation guidelines designed previously [6,9], we collected relevance judgments on Amazon Mechanical Turk for the 5,262 unique arguments in the top-5 pooling of the participants' runs. The most effective approach achieved an nDCG@5 of 0.81, the least effective 0.27, while the (argumentation-agnostic) baseline achieved an nDCG@5 of 0.77.

Task 2: Argument Retrieval for Comparative Questions. Five teams submitted eleven approaches to this task—all used the BM25F-based search engine ChatNoir [2] to access the ClueWeb12 and to retrieve an initial candidate ranking. For further re-ranking, models of different complexity were employed, basically in three steps: (1) represent documents and queries using language models, (2) identify arguments and comparative structures in documents, and (3) assess argument quality. Interestingly, only two approaches used query expansion techniques for retrieving the initial candidates.

Similar to the first task, we evaluated only the relevance of the retrieved documents. Using a top-5 pooling (10 submitted runs plus the additional ChatNoir baseline), a total of 1,783 unique results were judged by volunteers recruited internally. According to the evaluation, only one approach achieved a slightly better average nDCG@5 score than the ChatNoir baseline by using query expansion and taking credibility and argumentativeness into account in the re-ranking. The top-5 approaches (including the baseline) all had average nDCG@5 scores in the range of 0.55–0.58. Interestingly, the top-4 approaches relied on traditional feature engineering, while four out of the six lower-ranked approaches used deep learning-based language models. This difference in effectiveness could be caused by the absence of task-specific training data in the first lab edition. Using the relevance judgments created in the first lab may enable participants of the second lab edition to better train and fine-tune their (neural) re-ranking methods.

4 Conclusion

The main goal of Touché and its two shared tasks on argument retrieval for controversial and comparative questions is to establish a collaborative platform for researchers in the area of argument retrieval. By providing submission and evaluation tools as well as by organizing collaborative events such as workshops,

[9] https://trec.nist.gov/trec_eval/.

Touché aims to foster accumulating knowledge and developing new approaches in the field. All evaluation resources developed at Touché are shared freely, including search queries (topics), the assembled relevance judgments (qrels), and the participants' submitted ranked result lists (runs).

The evaluation of the approaches from 17 participating teams in the first Touché lab indicates that relatively basic, argumentation-agnostic baselines such as DirichletLM and BM25F-based retrieval are still almost as effective as the best approaches. Even though query expansion, argument quality assessment, or comparison features helped to (somewhat slightly) increase the overall retrieval effectiveness in the respective tasks' scenarios, there appears to be ample room for further improvement. More research on argument retrieval is thus well-justified.

In the second year of Touché, the participants are able to use the previously collected relevance judgments to develop and fine-tune new argument retrieval approaches, which may also allow for deploying state-of-the-art neural retrieval models. Moreover, we plan to have deeper judgment pools and to additionally evaluate argument quality dimensions, such as logical cogency and strength of support.

Acknowledgments. This work was partially supported by the DFG through the project "ACQuA: Answering Comparative Questions with Arguments" (grants BI 1544/7-1 and HA 5851/2-1) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999).

References

1. Ajjour, Y., Wachsmuth, H., Kiesel, J., Potthast, M., Hagen, M., Stein, B.: Data acquisition for argument search: the args.me corpus. In: Benzmüller, C., Stuckenschmidt, H. (eds.) KI 2019. LNCS (LNAI), vol. 11793, pp. 48–59. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30179-8_4
2. Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Elastic ChatNoir: search engine for the ClueWeb and the common crawl. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 820–824. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_83
3. Bondarenko, A., et al.: Overview of Touché 2020: argument retrieval. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 384–395. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_26
4. Braunstain, L., Kurland, O., Carmel, D., Szpektor, I., Shtok, A.: Supporting human answers for advice-seeking questions in CQA sites. In: Ferro, N., et al. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 129–141. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30671-1_10
5. Chernodub, A., et al.: TARGER: neural argument mining at your fingertips. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019 (demos), pp. 195–200, Association for Computational Linguistics (2019). URL https://doi.org/10.18653/v1/p19-3031

6. Gienapp, L., Stein, B., Hagen, M., Potthast, M.: Efficient pairwise annotation of argument quality. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pp. 5772–5781. Association for Computational Linguistics (2020). https://www.aclweb.org/anthology/2020.acl-main.511/
7. Gretz, S., et al.: A large-scale dataset for argument quality ranking: construction and analysis. In: Proceedings of The Thirty-Fourth Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, pp. 7805–7813. AAAI Press (2020). https://aaai.org/ojs/index.php/AAAI/article/view/6285
8. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002). http://doi.acm.org/10.1145/582415.582418
9. Potthast, M., et al.: Argument search: assessing argument relevance. In: Proceedings of the 42nd International Conference on Research and Development in Information Retrieval, SIGIR 2019, pp. 1117–1120. ACM (2019). https://doi.org/10.1145/3331184.3331327
10. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA integrated research architecture. Information Retrieval Evaluation in a Changing World. TIRS, vol. 41, pp. 123–160. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22948-1_5
11. Rafalak, M., Abramczuk, K., Wierzbicki, A.: Incredible: is (almost) all web content trustworthy? Analysis of psychological factors related to website credibility evaluation. In: Proceedings of the 23rd International World Wide Web Conference, WWW 2014, Companion Volume, pp. 1117–1122. ACM (2014). https://doi.org/10.1145/2567948.2578997
12. Robertson, S.E., Zaragoza, H., Taylor, M.J.: Simple BM25 extension to multiple weighted fields. In: Proceedings of the 13th International Conference on Information and Knowledge Management, CIKM 2004, pp. 42–49. ACM (2004). https://doi.org/10.1145/1031171.1031181
13. Schildwächter, M., Bondarenko, A., Zenker, J., Hagen, M., Biemann, C., Panchenko, A.: Answering comparative questions: better than ten-blue-links? In: Proceedings of the Conference on Human Information Interaction and Retrieval, CHIIR 2019, pp. 361–365. ACM (2019). https://doi.org/10.1145/3295750.3298916
14. Stab, C., et al.: ArgumenText: searching for arguments in heterogeneous sources. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2018, pp. 21–25. Association for Computational Linguistics (2018). https://www.aclweb.org/anthology/N18-5005
15. Toledo, A., et al.: Automatic argument quality assessment - new datasets and methods. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, pp. 5624–5634. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1564
16. Wachsmuth, H., et al.: Computational argumentation quality assessment in natural language. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 176–187. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/e17-1017
17. Wachsmuth, H., et al.: Building an argument search engine for the web. In: Proceedings of the Fourth Workshop on Argument Mining, ArgMining 2017, pp. 49–59. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/w17-5106

18. Yang, P., Fang, H., Lin, J.: Anserini: enabling the use of lucene for information retrieval research. In: Proceedings of the 40th International Conference on Research and Development in Information Retrieval, SIGIR 2017, pp. 1253–1256. ACM (2017). https://doi.org/10.1145/3077136.3080721
19. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th International Conference on Research and Development in Information Retrieval, SIGIR 2001, pp. 334–342. ACM (2001). https://doi.org/10.1145/383952.384019