Touché: First Shared Task on Argument Retrieval

Alexander Bondarenko[1], Matthias Hagen[1], Martin Potthast[2],
Henning Wachsmuth[3], Meriem Beloucif[4], Chris Biemann[4],
Alexander Panchenko[5], and Benno Stein[6]

[1] Martin-Luther-Universität Halle-Wittenberg, Germany
[2] Leipzig University, Germany
[3] Paderborn University, Germany
[4] Universität Hamburg, Germany
[5] Skolkovo Institute of Science and Technology, Russia
[6] Bauhaus-Universität Weimar, Germany

touche@webis.de

Abstract. Technologies for argument mining and argumentation processing are maturing continuously, giving rise to the idea of retrieving arguments in search scenarios. We introduce *Touché*, the first lab on *Argument Retrieval* featuring two subtasks: (1) the retrieval of arguments from a focused debate collection to support argumentative conversations, and (2) the retrieval of arguments from a generic web crawl to answer comparative questions with argumentative results. The goal of this lab is to perform an evaluation of various strategies to retrieve argumentative information from the web content. In this paper, we describe the setting of each subtask: the motivation, the data, and the evaluation methodology.

1 Introduction

Decision making processes, be it at the societal or at the personal level, eventually come to a point where one side will challenge the other with a *why*-question, i.e., a prompt to justify one's stance. In its most basic form, the answer to a why-question is a plain fact, but, more commonly, it requires a formulation of an argument, which is a justified claim.

The web is rife with documents comprising arguments, from news articles, blog posts, and discussion threads to advertisements and reviews of products and services. While the leading web search engines support the retrieval of plain facts often fairly well, hardly any support is currently provided for the retrieval of argumentative text, let alone the retrieval and ranking of individual arguments. This is particularly unfortunate in today's political and corresponding societal discourse, where well-reasoned argumentation on all kinds of controversial topics is of utmost importance. Yet, especially search results on such topics are often riddled with populism, conspiracy theories, and one-sidedness. This is not to say that these are not valid argumentative techniques, but rather that they arguably do not lead to the kind of information and insights one wants to support. Also at the personal level, users are very interested in (and often search for) the most relevant arguments that speak for or against a possible decision.

We propose a CLEF lab that aims for a better support of argument retrieval, making it possible to find "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 study abroad if I've never left my country before?"). In particular, the lab includes two subtasks, both of which are explained in detail below:

1. Argument retrieval from a focused debate collection to support argumentative conversations by providing justifications for the claims.
2. Argument retrieval from a generic web crawl to answer comparative questions with argumentative results and to support decision making.

With the proposed lab and its subtasks, our goal is to establish an understanding of how to evaluate argument retrieval processes as well as what kind of approaches effectively retrieve arguments that are beneficial for a variety of information needs and that are well-conceived in selected scenarios. An important component of the retrieval pipeline will be developing computational methods for the argument quality assessment (whether a given argument is a "strong" one), which is discussed in Section 4. This will not only allow for a better support of the argumentative information needs by search engines, but also, in the long run, may become an enabling technology for automatic open-domain agents that convincingly discuss and interact with humans.

2 Task Definition

The proposed lab will consist of the two subtasks to cover various information needs found in two different kinds of argumentative scenarios. The lab follows the classical TREC-style[7] evaluation methodology, where a dataset and a set of topics are provided to the participants. Each topic contains a search query and a detailed search scenario description. The task is to retrieve relevant documents satisfying conditions provided in the topic. Participants then submit their ranked retrieval results for each topic to be judged. Participating teams will be able to submit up to three runs with different approaches to be evaluated by expert assessors. We will make all the runs from this year's task available to the community — to have a basis for training models in light of a potential second edition of the lab as well as to enable reproducibility and independent research.

Task 1: Conversational Argument Retrieval

The first subtask is motivated by the support of users who search for arguments directly, e.g., by supporting their stance, or by aiding them in building a stance on topics of a general societal interest. Examples of such (usually controversial) topics are the abandonment of plastic bottles, animal experiments, immigration, and

[7] `https://trec.nist.gov/tracks.html`

abortion. Multiple online portals are centered around exactly such topics, such as Yahoo! Answers or Quora. Surprisingly, however, search engines nowadays do not provide any effective way to retrieve reliable arguments from these platforms, even though the search engines provide snippets for direct answers to factoid questions, among others. Presumably, the main reason is that search engines often do not grasp the argumentative nature of the underlying information need.

This subtask targets argumentative conversations. We will provide a focused crawl with content from online debate portals (idebate.org, debatepedia.org, debatewise.org) and from Reddit's ChangeMyView.[8] As a baseline retrieval model, we resort to the search engine args.me [10]. The lab participants will have to retrieve "strong" arguments (refer to Section 4 for a more detailed description of "strong") from the provided dataset for the 50 given topics, covering a wide range of controversial issues collected from the debate portals and ChangeMyView.

Task 2: Comparative Argument Retrieval

The second task is motivated by the support of users in personal decisions from everyday life where choices need to be made. In particular, the goal of the task is to find relevant arguments when comparing several objects with different options (e.g., "Is X better than Y for Z?"). Evidently, comparative information needs seem to be important: Question answering platforms such as Quora are filled with topics such as "How Python compares to PHP for web development?[9]" or "Is Germany better to live in compared to US?[10]". Still, in their current form, search engines such as Google and DuckDuckGo[11] do not provide much support for such comparative queries besides the "ten blue links". Retrieval of comparative information is thus eminent in web search and, according to [1], appears in about 10% of all search sessions. The retrieval topics for this task are based on real-world comparative questions that were submitted to commercial search engines and posted to question answering platforms.

The participants of this task will retrieve documents from a general web crawl (namely, ClueWeb12[12]) that help users to answer their comparative question. The task will be to identify documents that, ideally, comprise convincing argumentation for or against one or the other option underlying the comparative question. Two BM25F-based Elasticsearch systems, ChatNoir [2] and TARGER [5], will be available as baseline retrieval systems to participants that face problems or just want to avoid indexing the whole dataset on their side. Furthermore, TARGER's API can also be used to identify argumentative units in free text input. Additionally, we provide a baseline argument retrieval approach proposed by [3], which integrates the TARGER's API to capture "argumentativeness" in text documents. The basic idea applied in the approach is to axiomatically re-rank the top-50 results of BM25F for those topics that seem to be argumentative.

[8] https://www.reddit.com/r/changemyview

[9] https://www.quora.com/How-does-Python-compare-to-PHP-for-server-side-web-development

[10] https://www.quora.com/Is-Germany-better-to-live-in-compared-to-the-US

[11] https://duckduckgo.com

[12] https://lemurproject.org/clueweb12/

3 Data Description

For both subtasks, the topics have already been defined, ensuring that respective information is available in the focused crawl of debate portals and Reddit, and in the ClueWeb12, respectively. In total, we prepared 100 topics, 50 for each subtask. Each topic consists of a *title* representing either a search query or a choice question, a *description* providing a detailed definition of the search task, and a *narrative* accurately defining relevant documents.

Example topic for Task 1:

```
<title>
    climate change real
</title>
<description>
    You read an opinion piece on how climate change is a hoax
    and disagree. Now you are looking for arguments supporting
    the claim that climate change is in fact real.
</description>
<narrative>
    Relevant arguments will support the given stance that
    climate change is real or attack a hoax side's argument.
</narrative>
```

Example topic for Task 2:

```
<title>
    What are advantages and disadvantages of PHP over Python
    and vice versa?
</title>
<description>
    The user is looking for differences and similarities of PHP
    and Python and wants to know about scenarios that favor one
    over the other.
</description>
<narrative>
    Relevant documents may contain an overview of more than
    these two programming languages but must include both of
    them with an explicit comparison of these two.
</narrative>
```

The topics for the first subtask were chosen from the online debate portals having the largest number of the user-generated comments, and thus represent the matters of the highest societal interest. As for the second subtask, two types of web sources have been used to create the topics: Questions posted to Yahoo! Answers, as well as question queries submitted to either of two search engines, Yandex or Google. We thoroughly selected question queries which correspond to a choice problem, and where the answers should contain a sufficient number of pro and con arguments. We also ensured that relevant documents can be found in the provided search collection.

4 Evaluation

For the first subtask, the evaluation is based on the pooled top-k results of the participants' runs. For these, human assessors will label argumentative text passages or documents manually, both for their general topical relevance, and for argument quality dimensions, which have been found to be important for the evaluation of arguments [9]: 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., somewhat similar to the concept of "utility").

For the second subtask, the human assessors will judge, in addition to document relevance, whether a sufficient argumentative support is provided as defined by [4] and will evaluate the trustworthiness and credibility of the web documents as in [8]. Thus, a "strong" argument is defined to be the one fulfilling certain argument quality criteria such as logical cogency, rhetorical well-writtenness, contribution to a stance-building, level of support, and credibility.

For both subtasks, the performance and ordering of the submitted runs will be measured in traditional ranking-based ways with respect to relevance (e.g., graded relevance judgments nDCG [6], but ignoring repeated and near-duplicate entries). Moreover, we will include in the judgment the qualitative aspects of arguments. We plan to do assessment of the submitted runs via crowdsourcing. The study conducted by [7] shows that argument assessment is feasible via crowdsourcing; however, especially the argument quality-oriented aspects will be further developed over the course of prospective future editions of the lab: Our goal is to establish a widely accepted way of evaluating argument retrieval and to support targeted improvements of the retrieval technology developed by the lab participants.

5 Participants and Lab Organization

In essence, the first task is a general retrieval task with a focused collection. Participants of respective ad-hoc setups from previous years at TREC, CLEF, and NTCIR can participate with ease, either with their favorite system, or enriched with a pipeline taking argumentativeness into account. The second task will be centered around information needs expressed as questions, which is attractive to participants from previous question-oriented tracks/labs. Both subtasks ultimately aim to support the building of (retrieval) systems that can discuss and argue with human users. Hence, also the rapidly growing community around conversational search, with successful workshops at the recent SIGIR, WWW, IJCAI, and EMNLP (CAIR and SCAI), might be interested.

Argument mining has grown rapidly in the NLP community over the last couple of years, with flourishing publications and workshops at the premier conferences (e.g., the ArgMining workshop at ACL and EMNLP as well as the COMMA series of focused conferences). Using the baseline retrieval systems that we will provide for the lab, groups from the NLP community not familiar with setting up a search system can contribute runs (e.g., by re-ranking the baseline's initial result set regarding their ideas of what forms a good argument).

The proposed CLEF 2020 lab requires human assessors to annotate different aspects of the retrieved argumentative passages or documents. For both subtasks, the pooling depths of the participants' runs can be adjusted to fit the available assessment resources.

With regard to organization, we plan to release data and the baseline solutions early, which will allow for a quick start for participants. The received solutions and notebooks will be reviewed for quality by the organizing committee. Notebooks that do not meet rigorous quality standards (e.g., quality of method presentation, novelty factor, soundness of approach) may be limited in size, or entirely rejected. All participants will be encouraged to present a poster at the lab session. A selection of the contributions will be invited to present their systems orally.

Overall, the lab session will be divided into two major parts. The first part will be comprised of an introduction to all the participants' solutions, an invited keynote by a speaker from the area, and the best solution talks. The second part of the lab will include a second invited keynote, a poster session, and a plenary discussion to wrap-up the lab results and to discuss possible objectives for improvements and future directions. The expected length of the lab session at the conference is one day (≈ 8 hours) with an estimated amount of 20–25 participants, including 5 talks and 10 poster presentations plus the aforementioned keynotes and overview presentations.

6 Conclusion

We propose Touché, the first lab on Argument Retrieval at CLEF 2020 to support a growing interest in argumentation technologies. In particular, by providing data, baseline retrieval systems, and the evaluation of the participants' approaches, we encourage researchers to contribute ideas to be shared with the community. We suggest to integrate a notion of a "strong" argumentation, or argument quality, into a retrieval pipeline by exploiting definite metrics such as level of support, credibility, cogency, style, and contribution to a stance. Our lab is aimed at inducing a discussion and developing computational approaches to the underlying aspects of the argument retrieval from an application-driven perspective.

Acknowledgments

This work has been 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. Bailey, P., Chen, L., Grosenick, S., Jiang, L., Li, Y., Reinholdtsen, P., Salada, C., Wang, H., Wong, S.: User Task Understanding: A Web Search Engine Perspective. In: NII Shonan Meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems. pp. 44–52 (2012)
2. Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In: Proceedings of the 40th European Conference on IR Research (ECIR). pp. 820–824 (2018)
3. Bondarenko, A., Völske, M., Panchenko, A., Biemann, C., Stein, B., Hagen, M.: Webis at TREC 2018: Common Core Track. In: Voorhees, E., Ellis, A. (eds.) Proceedings of the 27th International Text Retrieval Conference (TREC) (2018)
4. Braunstain, L., Kurland, O., Carmel, D., Szpektor, I., Shtok, A.: Supporting Human Answers for Advice-Seeking Questions in CQA Sites. In: Proceedings of the 38th European Conference on IR Research, (ECIR). pp. 129–141 (2016)
5. Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., Panchenko, A.: TARGER: Neural Argument Mining at Your Fingertips. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 195–200 (2019)
6. Kanoulas, E., Aslam, J.A.: Empirical Justification of the Gain and Discount Function for nDCG. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, (CIKM). pp. 611–620 (2009)
7. Potthast, M., Gienapp, L., Euchner, F., Heilenkötter, N., Weidmann, N., Wachsmuth, H., Stein, B., Hagen, M.: Argument Search: Assessing Argument Relevance. In: Proceedings of the 42nd International ACM Conference on Research and Development in Information Retrieval (SIGIR). pp. 1117–1120 (2019)
8. 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), Companion Volume. pp. 1117–1122 (2014)
9. Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: 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 (2017)
10. Wachsmuth, H., Potthast, M., Al-Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., Stein, B.: Building an Argument Search Engine for the Web. In: Proceedings of the Fourth Workshop on Argument Mining (ArgMining) at EMNLP. pp. 49–59 (2017)