Overview of Touché 2022: Argument Retrieval
Extended Abstract

Alexander Bondarenko[1(✉)], Maik Fröbe[1], Johannes Kiesel[2], Shahbaz Syed[3], Timon Gurcke[4], Meriem Beloucif[5], Alexander Panchenko[6], Chris Biemann[5], Benno Stein[2], Henning Wachsmuth[4], Martin Potthast[3], and Matthias Hagen[1]

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

Abstract. The goal of the Touché lab on argument retrieval is to foster and support the development of technologies for argument mining and argument analysis. In the third edition of Touché, we organize three shared tasks: (a) argument retrieval for controversial topics, where participants retrieve a gist of arguments from a collection of online debates, (b) argument retrieval for comparative questions, where participants retrieve argumentative passages from a generic web crawl, and (c) image retrieval for arguments, where participants retrieve images from a focused web crawl that show support or opposition to some stance. In this paper, we briefly summarize the results of two years of organizing Touché and describe the planned setup for the third edition at CLEF 2022.

1 Introduction

Decision making and opinion formation are routine human tasks that often involve weighing pro and con arguments. Since the Web is full of argumentative texts on almost any topic, in principle, everybody has the chance to acquire knowledge to come to informed decisions or opinions by simply using a search engine. However, large amounts of the easily accessible arguments may be of low quality. For example, they may be irrelevant, contain incoherent logic, provide insufficient support, or use foul language. Such arguments should rather remain "invisible" in search results which implies several retrieval challenges—regardless of whether a query is about socially important topics or "only" about personal decisions. The challenges include assessing an argument's relevance to a query, deciding what is an argument's main "gist" in terms of the take-away, and estimating how well an implied stance is justified but also range to finding images that help to illustrate some stance. Still, today's popular web search engines do

M. Hagen et al. (Eds.): ECIR 2022, LNCS 13186, pp. 339–346, 2022.
https://doi.org/10.1007/978-3-030-99739-7_43

not really address these challenges and lack a sophisticated support for searchers in argument retrieval scenarios—a gap we aim to close with the Touché lab.[1]

In the spirit of the two successful Touché labs on argument retrieval at CLEF 2020 and 2021 [6,7], we propose a third lab edition to again bring together researchers from the fields of information retrieval and natural language processing who work on argumentation. At Touché 2022, we organize the following three shared tasks, the last of which being fully new to this edition:

1. Argumentative sentence retrieval from a focused collection (crawled from debate portals) to support argumentative conversations on controversial topics.
2. Argument retrieval from a large collection of text passages to support answering comparative questions in the scenario of personal decision making.
3. Image retrieval to corroborate and strengthen textual arguments and to provide a quick overview of public opinions on controversial topics.

As part of the previous Touché labs, we evaluated about 130 submissions from 44 teams; the majority submitted their software using the tira.io platform. Many of the submissions improved over the "official" argumentation-agnostic DirichletLM- and BM25-based baselines. In total, we manually assessed more than 11,000 argumentative texts and web documents for 200 search topics. All topics and judgments are publicly available at https://touche.webis.de.

While the first two Touché editions focused on retrieving complete arguments and documents, the third edition focuses on more refined problems. Three shared tasks explore whether argument retrieval can support decision making and opinion formation more directly by extracting the argumentative gist from documents, by classifying their stance as pro or con towards the issue in question, and by retrieving images that show support or opposition to some stance.

2 Task Definition

In the Touché lab, we follow the classic TREC-style[2] methodology: documents and topics are provided to the participants who then submit their ranked results (up to five runs) for every topic to be judged by human assessors. The third lab edition includes the three complementary tasks already sketched above and further detailed in the following: (1) argument retrieval for controversial questions, (2) argument retrieval for comparative questions, and (3) image retrieval for arguments. The unit of retrieval of our previous tasks were always entire documents, whereas now we focus on the retrieval of relevant argumentative sentences, passages, and images as well as their stance detection.

[1] 'touché' is commonly "used to acknowledge a hit in fencing or the success or appropriateness of an argument" [https://merriam-webster.com/dictionary/touche].

[2] https://trec.nist.gov/tracks.html

2.1 Task Description

Task 1: Argument Retrieval for Controversial Questions. Given a controversial topic and a collection of arguments, the task is to retrieve sentence pairs that represent one argument's gist (e.g., a claim in one sentence and a premise in the other), and to rank these pairs according to their relevance to the topic. The argument collection for Task 1 is the args.me corpus [1]. A pre-processed version of the args.me corpus with each argument split into its constituent sentences is provided and can be indexed easily by the participants.

The pairs retrieved by the participants will be evaluated by human assessors with respect to topical relevance and argument quality. As for quality, there are three key properties: (1) each sentence in the pair must be argumentative (e.g., a claim, a premise, or a conclusion), (2) the sentence pair must form a coherent text (e.g., sentences in a pair must not contradict each other), and (3) the sentence pair constitutes a short summary of a single argument (i.e., the major claim of an argument and the best premise supporting this claim are good candidates).

The participants may use a number of previously compiled resources to lower the entry barrier of this task. These include the document-level relevance and quality judgments from the previous Touché editions, and a sample of sentence pairs from the snippet generation framework of Alshomary et al. [3], enabling a basic understanding of the task and the evaluation during development. For the identification of claims and premises, the participants can use any existing argument tagging tool, such as the API[3] of TARGER [9] hosted on our own servers, or develop an own method if necessary.

Task 2: Argument Retrieval for Comparative Questions. Given a comparison search topic with two comparison objects and a collection of text passages, the task is to retrieve relevant argumentative passages for one or both objects, and to detect the passages' stances with respect to the two objects. The collection for Task 2 is a focused collection of 868,655 passages extracted from the ClueWeb12 for the 50 search topics of the task (cf. Sect. 2.2). Near-duplicates are already removed with CopyCat [12] to mitigate negative impacts [13,14].

The relevance of the top-k ranked passages of a system ($k \geq 5$ determined based on assessor load) will be assessed by human annotators ('not relevant', 'relevant', or 'highly relevant') along with the rhetorical quality [22] ('no arguments or low quality', 'average quality', or 'high quality'). Stance detection effectiveness will be evaluated in terms of the accuracy of distinguishing 'pro first compared object', 'pro second compared object', 'neutral', and 'no stance'.

The participants may use a number of previously compiled resources to lower the entry barrier of this task. These include the document-level relevance and argument quality judgments from the previous Touché editions as well as, for passage-level relevance judgments, a subset of MS MARCO [19] with comparative questions identified by our ALBERT-based [17] classifier (about 40,000 questions are comparative) [5]. Each comparative question in MS MARCO contains

[3] Also available as a Python library: https://pypi.org/project/targer-api/

10 text passages with relevance labels. For stance detection, a dataset comprising 950 comparative questions and answers extracted from Stack Exchange is provided [5]. For the identification of arguments in texts (e.g., claims and premises), the participants can use any existing argument tagging tool, such as the TARGER API hosted on our own servers, or develop their own tools.

Task 3: Image Retrieval for Arguments (New Task). Given a controversial topic and a collection of web documents with images, the task is to retrieve images that show support for each stance (pro/con the topic). The collection for Task 3 is a focused crawl of 10,000 images with the documents that contain them; for the retrieval, also the textual content of the web documents can be used.

A system's results should provide a searcher with a visual overview of public opinions on a controversial topic; we envision systems that juxtapose images for each stance. The approaches will be evaluated in terms of precision, namely by the ratio of relevant images among 20 retrieved images, 10 per stance.

Participants may use our available image-level relevance judgments [16]; The format is aligned with the format of the task's collection. Similar to the other Touché tasks, participants are free to use any additional existing tools and datasets or develop their own. Moreover, our goal is to collect a software suite for extracting various features—both for the images and web documents. Participants are encouraged to contribute Docker containers to this suite.

2.2 Search Topics

For the tasks on controversial questions (Task 1) and image retrieval (Task 3), we provide 50 search topics that represent a variety of debated societal matters. Each of these topics has a *title* in terms of a question on a controversial issue, a *description* specifying the particular search scenario, and a *narrative* that serves as a guideline for the human assessors:

```
<title> Should teachers get tenure? </title>
<description> A user has heard that some countries do give teachers
tenure and others don't. Interested in the reasoning for or against
tenure, the user searches for positive and negative arguments. [...]
</description>
<narrative> Highly relevant statements clearly focus on tenure for
teachers in schools or universities. Relevant statements consider tenure
more generally, not specifically for teachers, or [...] </narrative>
```

For the task on comparative questions (Task 2), we provide 50 search topics that describe scenarios of personal decision making. Each of these topics has a *title* in terms of a comparative question, *comparison objects* for the stance detection of the retrieved passages, a *description* specifying the particular search scenario, and a *narrative* that serves as a guideline for the assessors:

```
<title> Should I major in philosophy or psychology? </title>
<objects> major in philosophy, psychology </objects>
<description> A soon-to-be high-school graduate finds themselves at a
crossroad in their live. Based on their interests, majoring in philosophy
or in psychology are the potential options and the graduate is searching
for information about the differences and [...] </description>
<narrative> Relevant passages will overview one of the two majors in
terms of career prospects or developed new skills, or they will provide
reasons [...] </narrative>
```

3 Touché at CLEF 2021: Results and Findings

At Touché 2021, we received 36 registrations (compared to 28 registrations in the first year); aligned with the lab's fencing-related title, the participants were asked to select a real or fictional fencer or swordsman character (e.g., Zorro) as their team name upon registration. We received result submissions from 27 of the 36 registered teams (after 17 active teams in the first year) that resulted in 88 valid runs (after 41 in 2020; participants were allowed to submit up to 5 result rankings in both years). Touché aims to foster the reproducibility of submissions by asking participants to submit their approaches via the TIRA platform [20], which allows easy software submission and automatic evaluation.

Task 1: Argument Retrieval for Controversial Questions. In the first two Touché editions, Task 1 was stated as follows: given a question on a controversial topic, retrieve relevant and high-quality arguments from a focused crawl of online debate portals—the args.me corpus [1]. The submissions in 2021 [7] partly continued the trend of Touché 2020 [6] by deploying "traditional" retrieval models, however, with an increased focus on machine learning models (especially for query expansion and for argument quality assessment). Overall, there were two main trends in the participants' retrieval pipelines: (1) reproducing and fine-tuning approaches from the previous year by increasing their robustness, and (2) developing new, mostly neural approaches for argument retrieval by fine-tuning pre-trained models for the domain-specific search task at hand.

Like in the first year, combining "traditional" retrieval models with various query expansion methods and domain-specific re-ranking features remained a frequent choice for Task 1. Not really surprising—given its top effectiveness as the 2020 baseline—, DirichletLM was employed most often as the initial retrieval model, followed by BM25. For query expansion (e.g., with synonyms), most participating teams continued to use WordNet [11], however, Transformer-based approaches received increased attention [2]. Moreover, many approaches tried to use some form of argument quality estimation in the (re-)ranking.

The approaches in 2021 benefited from the relevance judgments collected at Touché in 2020. Many teams used them for general parameter optimization but also to evaluate intermediate results of their approaches, to select preprocessing methods, and to fine-tune or select the best configurations.

Task 2: Argument Retrieval for Comparative Questions. In the first two Touché editions, Task 2 was stated as follows: given a comparative question, retrieve documents from the ClueWeb12 that help to answer the comparative question. The participants' approaches submitted in 2021 all used the ChatNoir search engine [4] for an initial document retrieval, either by submitting the original topic titles as queries, or by applying query preprocessing (e.g., lemmatization and POS-tagging) and query expansion techniques (e.g., synonyms from WordNet [11], or generation based on word2vec [18] or sense2vec embeddings [21]). Most teams then applied a document "preprocessing" (e.g., removing HTML markup) before re-ranking the ChatNoir results with feature-based or neural classifiers trained on the Touché 2020 judgments (e.g., using argumentativeness, credibility, or comparativeness scores as features). The teams predicted document relevance labels by using a random forest classifier, XGBoost [8], LightGBM [15], or a fine-tuned BERT [10].

Overall, in both tasks, many more approaches submitted in 2021 could improve upon the argumentation-agnostic baselines (DirichletLM for Task 1 and BM25 for Task 2) than in the first year, indicating that progress was achieved.

4 Conclusion

At Touché, we continue our activities to establish a collaborative platform for researchers in the area of argument retrieval, and organize respective shared tasks for the third time. By providing submission and evaluation tools as well as by organizing collaborative events such as workshops, Touché aims to foster the accumulation of knowledge and development of new approaches in the field. All evaluation resources developed at Touché are shared freely, including search queries (topics), the assembled manual relevance and argument quality judgments (qrels), and the ranked result lists submitted by the participants (runs).

Acknowledgments. This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG) through the projects "ACQuA" and "ACQuA 2.0" (Answering Comparative Questions with Arguments; grants HA 5851/2-1, HA 5851/2-2, BI 1544/7-1, BI 1544/7-2) and "OASiS: Objective Argument Summarization in Search" (grant WA 4591/3-1), all part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999), and the German Ministry for Science and Education (BMBF) through the project "Shared Tasks as an Innovative Approach to Implement AI and Big Data-based Applications within Universities (SharKI)" (grant FKZ 16DHB4021). We are also grateful to Jan Heinrich Reimer for developing the TARGER Python library.

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. Akiki, C., Potthast, M.: Exploring argument retrieval with transformers. In: Working Notes Papers of the CLEF 2020 Evaluation Labs, vol. 2696 (2020), ISSN 1613–0073. http://ceur-ws.org/Vol-2696/
3. Alshomary, M., Düsterhus, N., Wachsmuth, H.: Extractive snippet generation for arguments. In: Proceedings of the 43nd International ACM Conference on Research and Development in Information Retrieval, SIGIR 2020, pp. 1969–1972, ACM (2020). https://doi.org/10.1145/3397271.3401186
4. 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
5. Bondarenko, A., Ajjour, Y., Dittmar, V., Homann, N., Braslavski, P., Hagen, M.: Towards understanding and answering comparative questions. In: Proceedings of the 15th ACM International Conference on Web Search and Data Mining, WSDM 2022. ACM (2022). https://doi.org/10.1145/3488560.3498534
6. Bondarenko, A., et al.: Overview of touché 2020: argument retrieval. In: Working Notes Papers of the CLEF 2020 Evaluation Labs, CEUR Workshop Proceedings, vol. 2696 (2020). https://doi.org/10.1007/978-3-030-58219-7_26
7. Bondarenko, A., et al.: Overview of touché 2021: argument retrieval. In: Candan, K.S., Ionescu, B., Goeuriot, L., Larsen, B., Müller, H., Joly, A., Maistro, M., Piroi, F., Faggioli, G., Ferro, N. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 450–467. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_28
8. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 785–794, ACM (2016). https://doi.org/10.1145/2939672.2939785
9. 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, pp. 195–200. ACL (2019). https://doi.org/10.18653/v1/p19-3031
10. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, pp. 4171–4186. ACL (2019). https://doi.org/10.18653/v1/n19-1423
11. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
12. Fröbe, M., Bevendorff, J., Gienapp, L., Völske, M., Stein, B., Potthast, M., Hagen, M.: CopyCat: near-duplicates within and between the ClueWeb and the common crawl. In: Proceedings of the 44th International ACM Conference on Research and Development in Information Retrieval, SIGIR 2021, pp. 2398–2404. ACM (2021). https://dl.acm.org/doi/10.1145/3404835.3463246
13. Fröbe, M., Bevendorff, J., Reimer, J., Potthast, M., Hagen, M.: Sampling Bias due to near-duplicates in learning to rank. In: Proceedings of the 43rd International ACM Conference on Research and Development in Information Retrieval, SIGIR 2020. ACM (2020). https://dl.acm.org/doi/10.1145/3397271.3401212
14. Fröbe, M., Bittner, J.P., Potthast, M., Hagen, M.: The effect of content-equivalent near-duplicates on the evaluation of search engines. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) ECIR 2020. LNCS, vol. 12036, pp. 12–19. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_2

15. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the Annual Conference on Neural Information Processing Systems, NeurIPS 2017, pp. 3146–3154 (2017). https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
16. Kiesel, J., Reichenbach, N., Stein, B., Potthast, M.: Image retrieval for arguments using stance-aware query expansion. In: Proceedings of the 8th Workshop on Argument Mining, ArgMining 2021 at EMNLP, pp. 36–45. ACL (2021)
17. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations. In: Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, OpenReview.net (2020). https://openreview.net/forum?id=H1eA7AEtvS
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the 1st International Conference on Learning Representations, ICLR 2013 (2013). http://arxiv.org/abs/1301.3781
19. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: a human generated machine reading comprehension dataset. In: Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 at NIPS, CEUR Workshop Proceedings, vol. 1773, CEUR-WS.org (2016). http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf
20. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA integrated research architecture. In: 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
21. Trask, A., Michalak, P., Liu, J.: sense2vec - a fast and accurate method for word sense disambiguation in neural word embeddings. CoRR abs/1511.06388 (2015). http://arxiv.org/abs/1511.06388
22. 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 2017, pp. 176–187 (2017). http://aclweb.org/anthology/E17-1017