Overview of Touché 2020: Argument Retrieval
Extended Abstract

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

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

touche@webis.de touche.webis.de

Abstract This paper is a condensed report on Touché: the first shared task on argument retrieval that was held at CLEF 2020. With the goal to create a collaborative platform for research in argument retrieval, we run two tasks: (1) supporting individuals in finding arguments on socially important topics and (2) supporting individuals with arguments on everyday personal decisions.

1 Introduction

Decision making and opinion formation processes are kind of routine tasks for many of us. Often, such opinion formation relates to a decision between two sides based on previous experience and knowledge, but it may also require accumulating new knowledge. With the wide-spread access to any kind of information on the web, everyone theoretically has the chance to acquire new knowledge and to form an informed opinion about any topic. In the process, be it on the level of socially important topics or "just" personal decisions, one of the at least two sides (i.e., decision options) will challenge the other with an appeal to justify its stance. In the simplest form, a justification might be simple facts or opinions, but more complex justifications often are based on argumentation: a complex relational aggregation of evidence and opinions, where one element is supported by the other.

Web resources such as blogs, community question answering websites, or social platforms contain an immense variety of opinions and argumentative texts—including many of biased, faked, or populist nature—which has motivated research on the development of high-quality argument retrieval. While standard web search engines support the retrieval of factual information fairly well, they hardly address the retrieval of argumentative texts specifically, let alone the retrieval and ranking of individual arguments

 CLEF 2020, 22–25 September 2020, Thessaloniki, Greece.

or opinions. In contrast, the argument search engine args.me [29] was developed to retrieve relevant arguments to a given controversial query. So far, however, it is limited to the document collections crawled from a few debating web portals. Other argument retrieval systems such as ArgumenText [26] and TARGER [8] take advantage of the large web document collection Common Crawl, but their ability to reliably retrieve arguments to support sides in a decision process is limited. The comparative argumentation machine CAM [25], a system for argument retrieval in comparative search, tries to support decision making in comparison scenarios based on billions of sentences from the Common Crawl but still lacks a proper ranking of diverse arguments.

To foster the research on a better support of argument retrieval, we organize the Touché lab at CLEF 2020—the first lab on argument retrieval [7].[1] The lab is a collaborative platform to develop retrieval approaches for decision support on a societal (e.g.,"Is climate change real and what to do?") and personal level (e.g.,"Should I buy real estate or rent, and why?") featuring two tasks:

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

Research on argument retrieval approaches will not only allow search engines to deliver more argumentative results for argumentative information needs (e.g., decision making in complex comparative search scenarios), but it will also be an important part of open-domain conversational agents that "discuss" controversial societal topics with humans—as showcased by IBM's Project Debater [17, 3].[2]

2 Previous Work

The input for argument retrieval can be a controversial topic, a question that compares two entities, or even a complete argument [31]. In the Touché lab, we address the first two types of information needs in two different shared tasks. Here, we summarize related work for both tasks.

2.1 Argument retrieval

Argument retrieval aims for delivering arguments to support users in taking a decision or persuading an audience with a specific point of view. An argument is usually modeled as a conclusion with supporting or attacking premises [29]. While a conclusion is a statement that can be accepted or rejected, a premise is a more grounded statement, e.g., a statistical evidence. The development of an argument search engine is faced with challenges that range from mining arguments from unstructured text to assessing

[1] The name of the lab is inspired by the usage of the term "touché" as an exclamation "used to admit that someone has made a good point against you in an argument or discussion." [https://dictionary.cambridge.org/dictionary/english/touche]

[2] https://www.research.ibm.com/artificial-intelligence/project-debater/

their relevance and quality [29]. Argument retrieval follows several paradigms that start from different sources and perform argument mining and retrieval tasks in different orders [1]. Wachsmuth et al. [29], e.g., extract arguments offline using heuristics that are tailored for online debate portals. The argument search engine args.me uses BM25F to rank arguments while giving conclusions more weight than premises. Levy et al. [15] uses distant-supervision to mine arguments offline for a set of topics from Wikipedia before ranking them. Stab et al. [26] retrieve documents from the Common Crawl[3] and then use a topic-dependent neural network to extract arguments from the retrieved documents. The two tasks in the Touché lab address the paradigms of Wachsmuth et al. [29] and Stab et al. [26] respectively.

Apart from its relevance to a topic, argument retrieval should rank arguments according to their quality. What makes a good argument has been studied since the time of Aristotle [2]. Recently, Wachsmuth et al. [28] categorized the different aspects of argument quality into a taxonomy that covers three dimensions: logic, rhetoric, and dialectic. Logic concerns the local structure of an argument, i.e, the conclusion and the premises and their relations. Rhetoric covers the effectiveness of the argument in persuading an audience with its conclusion. Dialectic addresses the relations of an argument to other arguments on the topic. For example, many attacking arguments make the argument vulnerable in a debate. The relevance of an argument to an input topic is categorized by Wachsmuth et al. [28] under dialectic quality. Researchers assess argument relevance by measuring its similarity to an input topic or incorporating its support/attack relations to other arguments. Potthast et al. [22] evaluate four standard retrieval models at ranking 437 arguments with regard to their quality. For argument quality, the researchers adopt three dimensions from Wachsmuth et al. [28]: logic, rhetoric, and dialectic. One of the main findings is that DirchletDM is better than BM25, DPH, and TF-IDF at ranking arguments. Gienapp et al. [10] extend this work by crowdsourcing a corpus of 1,271 arguments that are annotated in a pair-wise fashion with the same quality dimensions. The paper proposes a strategy that reduces costs by 93% by annotating only a subset of argument pairs. Wachsmuth et al. [30] create a graph of arguments by connecting two arguments if an argument uses another's conclusion as a premise. Later on, they exploit this structure to rank the arguments in the graph using PageRank [19]. This method is shown to outperform several baselines that utilize the content of the argument and its local structure (conclusion and premises). Dumani et al. [9] introduce a probabilistic framework that operates on semantically similar claims and premises. The framework utilizes the support/attack relations between the premises and claims clusters and the claims clusters and a query. The proposed framework is found to outperform BM25 in ranking arguments.

2.2 Comparative Argument Retrieval

User comparative information need was originally addressed in web search with the proposed simplistic interface, where the two compared objects would be separately typed in the search boxes on the left and right sides of the web interface [18, 27]. Additionally, opinion mining research has dealt with the identification of comparative

[3] http://commoncrawl.org

sentences and mining the user opinion (in favor or not) towards one or the other compared object in product reviews using Class Sequential Rule and SVM [12, 13, 14]. Recently, identification of the comparison preference ("winning" object) in comparative sentences has been addressed in open domain (not just product reviews) by applying feature-based and neural classifiers [21, 16]. This preference classification formed the basis of the comparative argumentation machine CAM [25], which is able to accept two compared objects and a comparison aspect as input, retrieves comparative sentences in favor of one or the other object using BM25, and clusters them in the for/against table to present to the user, but still lacks a proper ranking of diverse arguments.

3 Touché Task 1: Conversational Argument Retrieval

The goal of the Touché lab's first task is to provide assistance to users searching for good and relevant pro and con arguments on various societal topics (climate change, electric cars, etc.) while, for instance, being engaged in an argumentative conversation. A respective retrieval system may aid users in collecting evidence on issues of general societal interest and support them in forming their own opinion.

Several existing community question answering websites like Yahoo! Answers and Quora and also debating portals like debatewise.org or idebate.org are designed to accumulate opinions and arguments and to engage users in dialogues. General web search engines lack an effective solution to retrieve relevant arguments from these and other platforms beyond, for instance, simply returning complete longer threads. One reason probably is that the argumentative nature of the underlying discussions is ignored which results in general web search engines not really offering sufficient support during conversations or debates. This motivates the development of robust and effective approaches specifically focused on conversational argument retrieval.

3.1 Task Definition

The participants of Task 1 were asked to retrieve relevant arguments from a focused crawl of online debate portals for a given query on a controversial topic. Given the amount of argumentative texts readily available on online debate platforms, instead of extracting argumentative passages from unstructured text, the participants should build systems that retrieve items from a provided large collection of arguments covering a wide range of popular debate topics. For easy access to the document collection, we provided the openly accessible and flexible API of args.me,[4] also allowing participants to participate in the lab without having to index the collection on their end.

3.2 Data Description

Retrieval Topics. We have formulated 50 search scenarios on controversial issues in the form of TREC-style topics with a title (the query potentially issued by a user), a description (a short summary of the search context and information need), and a narrative

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

Table 1. Example Topic for Task 1: Conversational Argument Retrieval

Number	21
Title	Is human activity primarily responsible for global climate change?
Description	As the evidence that the climate is changing rapidly mounts, a user questions the common belief that climate change is anthropogenic and desires to know whether humans are the primary cause, or whether there are other causes.
Narrative	Highly relevant arguments include those that take a stance in favor of or opposed to climate change being anthropogenic and that offer valid reasons for either stance. Relevant arguments talk about human or non-human causes, but not about primary causes. Irrelevant arguments include ones that deny climate change.

(a definition of what constitutes relevant results for this topic, serving as a guideline for human assessors). An example topic is shown in Table 1. As topics, we selected those issues that have the largest number of user-generated comments on the debate portals, and thus probably having a high societal interest. Further, we ensured that relevant items for each topic are present in the provided document collection.

Document Collection. Task 1 is based on the args.me corpus [1] that is freely available for download[5] and also accessible via the mentioned args.me API. The corpus contains about 400,000 arguments crawled from four online debate portals: debatewise.org, idebate.org, debatepedia.org, and debate.org. Each argument in the corpus consists of a conclusion (claim) and one or more premises (reasons) supporting the conclusion.

3.3 Task Evaluation

In the first edition of the lab, we evaluate only the *relevance* of the retrieved documents (not the quality of the comprised arguments), given that the collection of manual judgments is a rather complex and time-consuming task. We collected the participants' results as classical TREC-style runs where, for each topic, the document IDs are returned in a ranked list ordered by descending relevance (i.e., the most relevant document should occur at Rank 1). The document pools for judgments were created with the TrecTools Python library [20][6] using a top-5 pooling strategy that resulted in 5,291 unique retrieval results to be judged.

The relevance judgments were collected on Amazon Mechanical Turk following previously designed annotation guidelines [10, 22]. We tasked the crowd workers to decide whether or not a given retrieved text is an argument, and to annotate the relevance of the item on a scale ranging from 1 (low relevance) to 5 (high relevance). Non-arguments were subsequently marked as spam and received a score of -2. Each retrieval result was separately annotated by five crowd workers, using majority vote as a decision rule. To further ensure the annotation quality, we recruited only workers for the task with an approval rate of at least 95%, and checked for occurrences of systematic spam.

[5] https://webis.de/data/args-me-corpus.html
[6] https://pypi.org/project/trectools/

We will evaluate the participants' approaches using nDCG [11] on the graded relevance judgments, and we will summarize the submitted approaches and report their results in the forthcoming complete lab overview [6].

4 Touché Task 2: Comparative Argument Retrieval

The goal of the Touché lab's second task is to support individuals' personal decisions in everyday life that can be expressed as a comparative information need ("Is X better than Y with respect to Z?") and that do not have a single "correct" answer. Such questions can, for instance, be found on community question answering (CQA) websites like Yahoo! Answers or Quora, or in discussions on Reddit, but are also submitted as queries to search engines. The search engines then often simply show content from CQA websites or some web document mentioning the query terms as a direct answer above the classic "ten blue links". However, a problem of such attempts at short direct answers is that CQA websites may not always provide a diverse and sufficient overview of all possible options with well-formulated arguments, nor will all underlying textual information be credible—a broader set of such issues recently was named as the dilemma of direct answers [24]. As a first step to work on technology to present several credible arguments and different angles in a search engine's potential direct comparative answers, we propose Task 2 on web-based comparative argument retrieval.

4.1 Task Definition

The participants of Task 2 were asked to retrieve and rank documents from the ClueWeb12[7] that help to answer a comparative question. Ideally, the retrieved documents contain convincing arguments for or against some of the possible options for a given comparison. Similar to Task 1, participation was possible without indexing the document collection on the participants' side since we provide easy access to the document collection through the BM25F-based ChatNoir search engine [4]—via a web-interface[8] and an API.[9] To identify arguments in texts, the participants were not restricted to any system; they could use own technology or any existing argument tagger of their choice. To lower the entry barriers for participants new to argument mining, we offered support for using the neural TARGER argument tagger [8] hosted on our own servers.

4.2 Data Description

Retrieval Topics. We selected 50 comparative questions from questions submitted to commercial search engines [5] or asked on question answering platforms, each covering some personal decision from everyday life. For every question, we have formulated a respective TREC-style topic with the question as the title, a description of the

[7] https://lemurproject.org/clueweb12/
[8] https://www.chatnoir.eu/
[9] https://www.chatnoir.eu/doc/

Table 2. Example Topic for Task 2: Comparative Argument Retrieval

Number	16
Title	Should I buy or rent?
Description	A person is planning to move out from their current small flat to start a family. Hoping that the new family will stay together in the new place for some longer time, the person is considering to even buy a new home and not just to rent it. However, this is kind of an important decision with many different angles to be considered: financial situation, the duties coming with owning a flat/house, potential happiness living in a property owned by someone else without any further (financial) responsibilities when major redos are needed, etc.
Narrative	Highly relevant documents contain various pros and cons for buying or renting a home. Particularly interesting could be checklists of what to favor in what situations. Documents containing definitions and "smaller" comparisons of buying or renting a property are relevant. Documents without any personal opinion/recommendation or pros/cons are not relevant.

searcher's possible context and information need, and a narrative describing what makes a result relevant (i.e., serving as a guideline for human assessors). An example topic is shown in Table 2. For each topic, we ensured that relevant documents are present in the ClueWeb12.

Document Collection. Task 2 is based on the ClueWeb12 document collection[10] crawled by the Language Technologies Institute at Carnegie Mellon University between February and May 2012 (733 million English web pages; 27.3TB uncompressed). Participants of Task 2 could index the ClueWeb12 on their own or could use the Elasticsearch-based ChatNoir API for a BM25F-based baseline retrieval.

4.3 Task Evaluation

Similar to Task 1, in the first edition of the lab, we evaluate only the relevance of the retrieved documents using a top-5 pooling strategy of the submitted participants' runs that resulted in 1,374 unique documents to be judged.

For the relevance judgments, we internally recruited seven grad and undergrad student volunteers, all with computer science background. We used a κ-test of five documents from five topics to "calibrate" the annotators' interpretations of the guidelines (i.e., the topics including the narratives) in follow-up discussions among the annotators. After the κ-test, the annotators judged disjoint subsets of the topics (each topic judged by one annotator only) and assigned one of three labels to a document: 0 (not relevant), 1 (relevant), or 2 (highly relevant).

We will evaluate the participants' approaches using nDCG [11] on the graded relevance judgments, and we will summarize the submitted approaches and report their results in the forthcoming complete lab overview [6].

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

5 Lab Overview and Statistics

A total of 28 teams registered, with a majority coming from Germany but also teams from the US, Europe, and Asia (17 from Germany, 2 from France, 2 from India, and 1 each from China, Italy, the Netherlands, Pakistan, Russia, Switzerland, and the US). As part of the registration, we asked the participants to choose as their team name a real or fictional fencer or swordsman character (e.g., Zorro)—aligned with the lab's fencing-related title.

From the 28 registered teams, 20 did submit results. To improve the reproducibility of the developed approaches, we asked the participants to use the TIRA platform [23] to also submit running software of their approaches. TIRA is an integrated cloud-based evaluation-as-a-service research architecture in which the participants have full administrative access to a virtual machine. By default, the virtual machines operate Ubuntu 18.04 with one CPU (Intel Xeon E5-2620), 4GB of RAM, and 16GB HDD, but we adjusted the resources to the participants' requirements when needed (e.g., one team asked for 24 GB of RAM, 5 CPUs, and 30 GB of HDD). Each virtual machine has standard software pre-installed (e.g., Docker and Python) to simplify the deployment of participants' approaches. After the deployment of an approach, the participants can create result submissions via the web UI of TIRA.

As an alternative to software submissions, we also allowed traditional run submissions but this option was only taken by 2 out of the 20 teams who submitted results. To allow a wide diversity of different approaches, we encouraged the teams to provide multiple solutions—asking the participants to prioritize runs/softwares when more than one was submitted. The runs needed to follow the standard TREC-style format.[11] Upon submission, we checked the validity and asked the participants to re-submit in case of problems, also offering our assistance. This resulted in 42 valid runs from 18 teams. From every team, the 5 runs with the highest priorities were used for the assessment pools.

To increase the reproducibility of participants' software submissions, TIRA follows a standard pipeline. To create a run submission from a participating team's software, the respective virtual machine is shut down, disconnected from the internet, powered on, and the datasets for the respective task are mounted in a sandbox mode. The interruption of the internet connection ensures that the participants' software works without external web services that may disappear or get incompatible in the future, which could reduce the reproducibility. However, we enabled two exceptions from the interruption of the internet connection for all participants: the APIs of ChatNoir and args.me were available, even in the sandbox mode. Additionally, we allowed external web services based on the participants' requirements, but only one team additionally asked to access the web of Trust API.[12] We will archive all the virtual machines that participants have used to make submissions to the Touché lab. This way, all submitted pieces of software can be re-evaluated or applied to new datasets as long as the APIs of the used web services remain available.

[11] Also described on the lab website: https://touche.webis.de

[12] https://www.mywot.com/developers

6 Summary and Outlook

In this paper, we have briefly reported on the Touché lab at CLEF 2020—the first shared task on argument retrieval. Touché features two tasks: (1) conversational argument retrieval to support argumentation on socially important problems in dialogue or debate scenarios, and (2) comparative argument retrieval to support decision making on a personal level. From 28 registered teams, 18 submitted at least one valid run. The respective evaluation results and an overview of the developed approaches will be part of the forthcoming complete lab overview [6].

For the next iteration of the Touché lab, we plan to have deeper judgment pools and to also evaluate an argument's quality dimensions like logical cogency or strength of support.

Acknowledgments

This work was 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).

Bibliography

[1] Ajjour, Y., Wachsmuth, H., Kiesel, J., Potthast, M., Hagen, M., Stein, B.: Data Acquisition for Argument Search: The args.me Corpus. In: Proceedings of the 42nd German Conference AI, KI 2019. Lecture Notes in Computer Science, vol. 11793, pp. 48–59. Springer (2019), https://doi.org/10.1007/978-3-030-30179-8_4

[2] Aristotle, Kennedy, G.A.: On Rhetoric: A Theory of Civic Discourse. Oxford: Oxford University Press (2006)

[3] Bar-Haim, R., Krieger, D., Toledo-Ronen, O., Edelstein, L., Bilu, Y., Halfon, A., Katz, Y., Menczel, A., Aharonov, R., Slonim, N.: From Surrogacy to Adoption; From Bitcoin to Cryptocurrency: Debate Topic Expansion. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019. pp. 977–990. Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/p19-1094

[4] 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 2018. Lecture Notes in Computer Science, vol. 10772, pp. 820–824. Springer (2018), https://doi.org/10.1007/978-3-319-76941-7_83

[5] Bondarenko, A., Braslavski, P., Völske, M., Aly, R., Fröbe, M., Panchenko, A., Biemann, C., Stein, B., Hagen, M.: Comparative Web Search Questions. In: Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM 2020. pp. 52–60. ACM (2020), https://doi.org/10.1145/3336191.3371848

[6] Bondarenko, A., Fröbe, M., Beloucif, M., Gienapp, L., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2020: Conversational and Comparative Argument Retrieval. In: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. (to appear). CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2020)

[7] Bondarenko, A., Hagen, M., Potthast, M., Wachsmuth, H., Beloucif, M., Biemann, C., Panchenko, A., Stein, B.: Touché: First Shared Task on Argument Retrieval. In:

Proceedings of the 42nd European Conference on IR Research, ECIR 2020. Lecture Notes in Computer Science, vol. 12036, pp. 517–523. Springer (2020), https://doi.org/10.1007/978-3-030-45442-5_67
[8] Chernodub, A.N., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., Panchenko, A.: TARGER: Neural Argument Mining at Your Fingertips. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019. pp. 195–200. Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/p19-3031
[9] Dumani, L., Neumann, P.J., Schenkel, R.: A Framework for Argument Retrieval - Ranking Argument Clusters by Frequency and Specificity. In: In Proceedings of the 42nd European Conference on IR Research, ECIR 2020. Lecture Notes in Computer Science, vol. 12035, pp. 431–445. Springer (2020), https://doi.org/10.1007/978-3-030-45439-5_29
[10] 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/
[11] 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
[12] Jindal, N., Liu, B.: Identifying Comparative Sentences in Text Documents. In: Proceedings of the 29th Annual International Conference on Research and Development in Information Retrieval, SIGIR 2006. pp. 244–251. ACM (2006), https://doi.org/10.1145/1148170.1148215
[13] Jindal, N., Liu, B.: Mining Comparative Sentences and Relations. In: Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference, AAAI 2006. pp. 1331–1336. AAAI Press (2006), http://www.aaai.org/Library/AAAI/2006/aaai06-209.php
[14] Kessler, W., Kuhn, J.: A Corpus of Comparisons in Product Reviews. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014. pp. 2242–2248. European Language Resources Association (ELRA) (2014), http://www.lrec-conf.org/proceedings/lrec2014/summaries/1001.html
[15] Levy, R., Bogin, B., Gretz, S., Aharonov, R., Slonim, N.: Towards an argumentative content search engine using weak supervision. In: Bender, E.M., Derczynski, L., Isabelle, P. (eds.) Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018. pp. 2066–2081. Association for Computational Linguistics (2018), https://www.aclweb.org/anthology/C18-1176/
[16] Ma, N., Mazumder, S., Wang, H., Liu, B.: Entity-Aware Dependency-Based Deep Graph Attention Network for Comparative Preference Classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. pp. 5782–5788. Association for Computational Linguistics (2020), https://www.aclweb.org/anthology/2020.acl-main.512/
[17] Mass, Y., Shechtman, S., Mordechay, M., Hoory, R., Shalom, O.S., Lev, G., Konopnicki, D.: Word Emphasis Prediction for Expressive Text to Speech. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association, Interspeech 2018. pp. 2868–2872. ISCA (2018), https://doi.org/10.21437/Interspeech.2018-1159
[18] Nadamoto, A., Tanaka, K.: A Comparative Web Browser (CWB) for Browsing and Comparing Web Pages. In: Proceedings of the 12th International World Wide Web Conference, WWW 2003. pp. 727–735. ACM (2003), https://doi.org/10.1145/775152.775254
[19] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web (1998)

[20] Palotti, J.R.M., Scells, H., Zuccon, G.: TrecTools: an Open-source Python Library for Information Retrieval Practitioners Involved in TREC-like Campaigns. In: Proceedings of the 42nd International Conference on Research and Development in Information Retrieval, SIGIR 2019. pp. 1325–1328. ACM (2019), https://doi.org/10.1145/3331184.3331399
[21] Panchenko, A., Bondarenko, A., Franzek, M., Hagen, M., Biemann, C.: Categorizing Comparative Sentences. In: Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019. pp. 136–145. Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/w19-4516
[22] 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 Conference on Research and Development in Information Retrieval, SIGIR 2019. pp. 1117–1120. ACM (2019), https://doi.org/10.1145/3331184.3331327
[23] Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, The Information Retrieval Series, vol. 41, pp. 123–160. Springer (2019), https://doi.org/10.1007/978-3-030-22948-1_5
[24] Potthast, M., Hagen, M., Stein, B.: The Dilemma of the Direct Answer. SIGIR Forum **54**(1) (Jun 2020), http://sigir.org/forum/issues/june-2020/
[25] 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
[26] Stab, C., Daxenberger, J., Stahlhut, C., Miller, T., Schiller, B., Tauchmann, C., Eger, S., Gurevych, I.: ArgumenText: Searching for Arguments in Heterogeneous Sources. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2018. pp. 21–25. Association for Computational Linguistics (2018), https://doi.org/10.18653/v1/n18-5005
[27] Sun, J., Wang, X., Shen, D., Zeng, H., Chen, Z.: CWS: A Comparative Web Search System. In: Proceedings of the 15th International Conference on World Wide Web, WWW 2006. pp. 467–476. ACM (2006), https://doi.org/10.1145/1135777.1135846
[28] Wachsmuth, H., Naderi, N., Habernal, I., Hou, Y., Hirst, G., Gurevych, I., Stein, B.: Argumentation Quality Assessment: Theory vs. Practice. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017. pp. 250–255. Association for Computational Linguistics (2017), https://doi.org/10.18653/v1/P17-2039
[29] Wachsmuth, H., Potthast, M., Khatib, K.A., 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 4th Workshop on Argument Mining, ArgMining@EMNLP 2017. pp. 49–59. Association for Computational Linguistics (2017), https://doi.org/10.18653/v1/w17-5106
[30] Wachsmuth, H., Stein, B., Ajjour, Y.: "PageRank" for Argument Relevance. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017. pp. 1117–1127. Association for Computational Linguistics (2017), https://doi.org/10.18653/v1/e17-1105
[31] Wachsmuth, H., Syed, S., Stein, B.: Retrieval of the Best Counterargument without Prior Topic Knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018. pp. 241–251. Association for Computational Linguistics (2018), https://www.aclweb.org/anthology/P18-1023/