Overview of Touché 2021: Argument Retrieval

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. This paper is a condensed report on the second year of the Touché shared task on argument retrieval held at CLEF 2021. With the goal to provide a collaborative platform for researchers, we organized two tasks: (1) supporting individuals in finding arguments on controversial topics of social importance and (2) supporting individuals with arguments in personal everyday comparison situations.

Keywords: Argument retrieval for controversial questions · Argument retrieval for comparative questions · Shared task

1 Introduction

Informed decision making and opinion formation are natural routine tasks. Generally, both of these tasks often involve weighing two or more options. Any choice to be made may be based on personal prior knowledge and experience, but they may also often require searching and processing new knowledge. With the ubiquitous access to various kinds of information on the web—from facts over opinions and anecdotes to arguments—everybody has the chance to acquire knowledge for decision making or opinion formation on almost any topic. However, large amounts of easily accessible information imply challenges such as the need to assess their relevance to the specific topic of interest and to estimate how well an implied stance is justified; no matter whether it is about topics of social importance or "just" about personal decisions. In the simplest form, such a

K. S. Candan et al. (Eds.): CLEF 2021, LNCS 12880, pp. 450–467, 2021.
https://doi.org/10.1007/978-3-030-85251-1_28

justification might be a collection of basic facts and opinions. More complex justifications are often grounded in argumentation, though; for instance, a complex relational aggregation of assertions and evidence pro or con either side, where different assertions or evidential statements support or refute each other.

Furthermore, while web resources such as blogs, community question answering sites, news articles, or social platforms contain an immense variety of opinions and argumentative texts, a notable proportion of these may be of biased, faked, or populist nature. This has motivated argument retrieval research to focus not only on the relevance of arguments, but also on the aspect of their quality. While conventional web search engines support the retrieval of factual information fairly well, they hardly address the deeper analysis and processing of argumentative texts, in terms of mining argument units from these texts, assessing the quality of the arguments, or classifying their stance. To address this, the argument search engine args.me [51] was developed to retrieve arguments relevant to a given controversial topic and to account for the pro or con stance of individual arguments in the result presentation. So far, however, it is limited to a document collection crawled from a few online debate portals, and largely disregards quality aspects. Other argument retrieval systems such as ArgumenText [45] and TARGER [13] 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 [44], a system for argument retrieval in comparative search, tries to support decision making in comparison scenarios based on billions of individual sentences from the Common Crawl. Still, it lacks a proper ranking of diverse longer argumentative texts.

To foster research on argument retrieval and to establish more collaboration and exchange of ideas and datasets among researchers, we organized the second Touché lab on argument retrieval at CLEF 2021 [8,9].[1] Touché is a collaborative platform[2] to develop and share retrieval approaches that aim to support decisions at a societal level (e.g., "Should hate speech be penalized more, and why?") and at a personal level (e.g., "Should I major in philosophy or psychology, and why?"), respectively. The second year of Touché featured two tasks:

1. Argument retrieval for *controversial* questions from a focused collection of debates to support opinion formation on topics of social importance.
2. Argument retrieval for *comparative* questions from a generic web crawl to support informed decision making.

Approaches to these two tasks, which do not only consider the relevance of arguments but also facets of argumentative quality, will help search engines to deliver more accurate argumentative results. Additionally, they will also be an important part of open-domain conversational agents that "discuss"

[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://touche.webis.de/.

controversial societal topics with humans—as showcased by IBM's Project Debater [4,5,32].[3]

The teams that participated in the second year of Touché were able to use the topics and relevance judgments from the first year to develop their approaches. Many trained and optimized learning-based rankers as part of their retrieval pipelines and employed a large variety of pre-processing methods (e.g., stemming, duplicate removal, query expansion), argument quality features, or comparative features (e.g., credibility, part-of-speech tags). In this paper, we report the results and briefly describe the most effective participants' retrieval approaches submitted at Touché 2021; a more comprehensive overview of each approach will be covered in the forthcoming extended overview [9].

2 Previous Work

Queries in argument retrieval often are phrases that describe a controversial topic, questions that ask to compare two options, or even complete arguments themselves [53]. In the Touché lab, we address the first two types in two different shared tasks. Here, we briefly summarize the related work for both tasks.

2.1 Argument Retrieval

Argument retrieval aims for delivering arguments to support users in making a decision or to help persuading an audience of a specific point of view. An argument is usually modeled as a conclusion with supporting or attacking premises [51]. 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 their relevance and quality [51]. Argument retrieval follows several paradigms that start from different sources and perform argument mining and retrieval tasks in different orders [1]. Wachsmuth et al. [51], for instance, extract arguments offline using heuristics that are tailored to online debate portals. Their argument search engine args.me uses BM25F to rank the indexed arguments while giving conclusions more weight than premises. Also Levy et al. [29] use distant supervision to mine arguments offline for a set of topics from Wikipedia before ranking them. Following a different paradigm, Stab et al. [45] retrieve documents from the Common Crawl[4] in an online fashion (no prior offline argument mining) and use a topic-dependent neural network to extract arguments from the retrieved documents at query time. With the two Touché tasks, we address the paradigms of Wachsmuth et al. [51] (Task 1) and Stab et al. [45] (Task 2), respectively.

Argument retrieval should rank arguments according to their topical relevance but also to their quality. What makes a good argument has been studied

[3] https://www.research.ibm.com/artificial-intelligence/project-debater/.
[4] http://commoncrawl.org.

since the time of Aristotle [3]. Recently, Wachsmuth et al. [48] 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, an argument that has many attacking premises might be rather vulnerable in a debate. The relevance of an argument to a query's topic is categorized by Wachsmuth et al. [48] under dialectic quality.

Researchers assess argument relevance by measuring an argument's similarity to a query's topic or incorporating its support/attack relations to other arguments. Potthast et al. [40] evaluate four standard retrieval models at ranking arguments with regard to the quality dimensions: relevance, logic, rhetoric, and dialectic. One of the main findings is that DirichletLM is better at ranking arguments than BM25, DPH, and TF-IDF. Gienapp et al. [21] extend this work by proposing a pairwise strategy that reduces the costs of crowdsourcing argument retrieval annotations in a pairwise fashion by 93% (i.e., annotating only a small subset of argument pairs).

Wachsmuth et al. [52] create a graph of arguments by connecting two arguments when one uses the other's conclusion as a premise. Later on, they exploit this structure to rank the arguments in the graph using PageRank scores [37]. This method is shown to outperform several baselines that only consider the content of the argument and its local structure (conclusion and premises). Dumani et al. [15] introduce a probabilistic framework that operates on semantically similar claims and premises. The framework utilizes support/attack relations between clusters of premises and claims and between clusters of claims and a query. It is found to outperform BM25 in ranking arguments. Later, Dumani et al. [16] also proposed an extension of the framework to include the quality of a premise as a probability by using the fraction of premises which are worse with regard to the three quality dimensions cogency, reasonableness, and effectiveness. Using a pairwise quality estimator trained on the Dagstuhl-15512 ArgQuality Corpus [50], their probabilistic framework with the argument quality component outperformed the one without it on the 50 Task 1 topics of Touché 2020.

2.2 Retrieval for Comparisons

Comparative information needs in web search have first been addressed by basic interfaces where two to-be-compared products are entered separately in a left and a right search box [34,46]. Comparative sentences are then identified and mined from product reviews in favor or against one or the other to-be-compared entity using opinion mining approaches [23,24,26]. Recently, the identification of the comparison preference (the "winning" entity) in comparative sentences has been tackled in a more broad domain (not just product reviews) by applying feature-based and neural classifiers [31,39]. Such preference classification forms the basis of the comparative argumentation machine CAM [44] that takes two entities and some comparison aspect(s) as input, retrieves comparative sentences in favor of

one or the other entity using BM25, and then classifies their preference for a final merged result table presentation. A proper argument ranking, however, is still missing in CAM. Chekalina et al. [11] later extend the system to accept comparative questions as input and to return a natural language answer to the user. A comparative question is parsed by identifying the comparison objects, aspect(s), and predicate. The system's answer is either generated directly based on Transformers [14] or by retrieval from an index of comparative sentences.

3 Lab Overview and Statistics

The second edition of Touché received 36 registrations (compared to 28 registrations in the first year), with a majority coming from Germany and Italy, but also from the Americas, Europe, Africa, and Asia (16 from Germany, 10 from Italy, 2 from the United States and Mexico, and 1 each from Canada, India, the Netherlands, Nigeria, the Russian Federation, and Tunisia). Aligned with the lab's fencing-related title, the participants were asked to select a real or fictional swordsman character (e.g., Zorro) as their team name upon registration.

We received result submissions from 27 of the 36 registered teams (up from 20 submissions in the first year). As in the previous edition of Touché, we paid attention to foster the reproducibility of the developed approaches by using the TIRA platform [41]. Upon registration, each team received an invitation to TIRA to deploy actual software implementations of their approaches. TIRA is an integrated cloud-based evaluation-as-a-service research architecture on which participants can install their software on a dedicated virtual machine. By default, the virtual machines operate the server version of Ubuntu 20.04 with one CPU (Intel Xeon E5-2620), 4 GB of RAM, and 16 GB HDD, but we adjusted the resources to the participants' requirements when needed (e.g., one team asked for 30 GB of RAM, 3 CPUs, and 30 GB of HDD). The participants had full administrative access to their virtual machines. Still, we pre-installed the latest versions of reasonable standard software (e.g., Docker and Python) to simplify the deployment of the approaches.

Using TIRA, the teams could create result submissions via a click in the web UI that then initiated the following pipeline: the respective virtual machine is shut down, disconnected from the internet, and powered on again in a sandbox mode, mounting the test datasets for the respective tasks, and running a team's deployed approach. The interruption of the internet connection ensures that the participants' software works without external web services that may disappear or become incompatible—possible causes of reproducibility issues—but it also means that downloading additional external code or models during the execution was not possible. We offered our support when this connection interruption caused problems during the deployment, for instance, with spaCy that tries to download models if they are not already available on the machine, or with PyTerrier that, in its default configuration, checks for online updates. To simplify participation of teams that do not want to develop a fully-fledged retrieval pipeline on their end, we enabled two exceptions from the interruption

of the internet connection for all participants: the APIs of args.me and ChatNoir were available even in the sandbox mode to allow accessing a baseline system for each of the tasks. The virtual machines that the participants used for their submissions will be archived such that the respective systems can be re-evaluated or applied to new datasets as long as the APIs of ChatNoir and args.me remain available—that are both maintained by us.

In cases where a software submission in TIRA was not possible, the participants could submit just run files. Overall, 5 of the 27 teams submitted traditional run files instead of software in TIRA. Per task, we allowed each team to submit up to 5 runs that should follow the standard TREC-style format.[5] We checked the validity of all submitted run files, asking participants to resubmit their run files (or software) if there were any validity issues—again, also offering our support in case of problems. All 27 teams submitted valid runs, resulting in 90 valid runs (doubling the 42 result submissions that we received in the first year).

4 Task 1: Argument Retrieval for Controversial Questions

The goal of the Touché 2021 lab's first task was to advance technologies that support individuals in forming opinions on socially important controversial topics such as: "Should hate speech be penalized more?". For such topics, the task was to retrieve relevant and high-quality argumentative texts from the args.me corpus [1], a focused crawl of online debate portals. In this scenario, relevant arguments should help users to form an opinion on the topic and to find arguments that are potentially useful in debates or discussions.

The results of last year's Task 1 participants indicated that improving upon "classic" argument-agnostic baseline retrieval models (such as BM25 and DirichletLM) in the ranking of arguments from a focused crawl is difficult, but, at the same time, the results of these baselines still left some room for improvements. Also, the detection of the degree of argumentativeness and the assessment of the quality of an argument were not "solved" in the first year, but identified as potentially interesting contributions of submissions to the task's second iteration.

4.1 Task Definition

Given a controversial topic formulated as a question, approaches to Task 1 needed to retrieve relevant and high-quality arguments from the args.me corpus, which covers a wide range of timely controversial topics. To enable approaches that leverage training and fine-tuning, the topics and relevance judgments from the 2020 edition of Task 1 were provided.

[5] The expected format of submissions was also described at https://touche.webis.de.

Table 1. Example topic for Task 1: Argument Retrieval for Controversial Questions.

Number	89
Title	Should hate speech be penalized more?
Description	Given the increasing amount of online hate speech, a user questions the necessity and legitimacy of taking legislative action to punish or inhibit hate speech.
Narrative	Highly relevant arguments include those that take a stance in favor of or opposed to stronger legislation and penalization of hate speech and that offer valid reasons for either stance. Relevant arguments talk about the prevalence and impact of hate speech, but may not mention legal aspects. Irrelevant arguments are the ones that are concerned with offensive language that is not directed towards a group or individuals on the basis of their membership in the group.

4.2 Data Description

Topics. We formulated 50 new search questions on controversial topics. Each topic consisted of (a) a title in form of a question that a user might submit as a query to a search engine, (b) a description that summarizes the particular information need and search scenario, and (c) a narrative that guides the assessors in recognizing relevant results (an example topic is given in Table 1). We carefully designed the topics by clustering the debate titles in the args.me corpus, formulating questions for a balanced mix of frequent and niche topics—manually ensuring that at least some relevant arguments are contained in the args.me corpus for each topic.

Document Collection. The document collection for Task 1 was the args.me corpus [1], which is freely available for download[6] and also accessible via the args.me API.[7] The corpus contains about 400,000 structured arguments (from debatewise.org, idebate.org, debatepedia.org, and debate.org), each with a conclusion (claim) and one or more supporting or attacking premises (reasons).

4.3 Submitted Approaches

Twenty-one participating teams submitted at least one valid run to Task 1. The submissions partly continued the trend of Touché 2020 [7] by deploying "classical" retrieval models, however with an increased focus on machine learning models (especially for query expansion and for assessing argument quality). Overall, we observed two kinds of contributions: (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 pretrained models for the domain-specific search task at hand.

[6] https://webis.de/data.html#args-me-corpus.
[7] https://www.args.me/api-en.html.

Like in the first year, combining "classical" retrieval models with various query expansion methods and domain-specific re-ranking features remained a frequent choice of approaches to Task 1. Not really surprising—given last year's baseline results—DirichletLM was employed most often as the initial retrieval model, followed by BM25. For query expansion, most participating teams continued to leverage WordNet [17]. However, transformer-based approaches received increased attention, such as query hallucination, which was successfully used by Akiki and Potthast [2] in the previous Touché lab. Similarly, utilizing deep semantic phrase embeddings to calculate the semantic similarity between a query and possible result documents gained widespread adoption. Moreover, many approaches tried to use some form of argument quality estimation as one of their features for ranking or re-ranking.

This year's approaches benefited from the judgments released for Touché in 2020. Many teams used them for general parameter optimization but also to evaluate intermediate results of their approaches and to fine-tune or select the best configurations. For instance, comparing different kinds of pre-processing methods based on the available judgments from last year received much attention (e.g., stopword lists, stemming algorithms, or duplicate removal).

4.4 Task Evaluation

The teams' result rankings should be formatted in the "standard" TREC format where document IDs are sorted by descending relevance score for each search topic (i.e., the most relevant argument/document occurs at Rank 1). Prior to creating the assessment pools, we ran a near-duplicate detection for all submitted runs using the CopyCat framework [18], since near-duplicates might impact evaluation results [19,20]. The framework found only 1.1% of the arguments in the top-5 results to be near-duplicates (mostly due to debate portal users reusing their arguments in multiple debate threads). We created duplicate-free versions of each result list by removing the documents for which a higher-ranked document is a near-duplicate; in such cases, the next ranked non-near-duplicate then just moved up the ranked list. The top-5 results of the original and the deduplicated runs then formed the judgment pool—created with TrecTools [38]—resulting in 3,711 unique documents that were manually assessed with respect to their relevance and argumentative quality.

For the assessment, we used the Doccano tool [35] and followed previously suggested annotation guidelines [21,40]. Our eight graduate and undergraduate student volunteers (all with a computer science background) assessed each argument's relevance to the given topic with four labels (0: not relevant, 1: relevant, 2: highly relevant, or -2: spam) and the argument's rhetorical quality [50] with three labels (0: low quality, 1: sufficient quality, and 2: high quality). To calibrate the annotators' interpretations of the guidelines (i.e., the topics including the narratives and instructions on argument quality), we performed an initial κ-test in which each annotator had to label the same 15 documents from three topics (5 documents from each topic). The observed Fleiss' κ values of 0.50 for argument relevance (moderate agreement) and of 0.39 for argument quality (fair

Table 2. Results for Task 1: Argument Retrieval for Controversial Questions. The left part (a) shows the evaluation results of a team's best run according to the results' relevance, while the right part (b) shows the best runs according to the results' quality. An asterisk (⋆) indicates that the runs with the best relevance and the best quality differ for a team. The baseline DirichletLM ranking is shown in bold.

(a) Best relevance score per team

	nDCG@5	
Team	**Rel.**	**Qual.**
Elrond⋆	0.720	0.809
Pippin Took⋆	0.705	0.798
Robin Hood⋆	0.691	0.756
Asterix⋆	0.681	0.802
Dread Pirate Roberts⋆	0.678	0.804
Skeletor⋆	0.667	0.815
Luke Skywalker	0.662	0.808
Shanks⋆	0.658	0.790
Heimdall⋆	0.648	0.833
Athos	0.637	0.802
Goemon Ishikawa	0.635	0.812
Jean Pierre Polnareff	0.633	0.802
Swordsman	**0.626**	**0.796**
Yeagerists	0.625	0.810
Hua Mulan⋆	0.620	0.789
Macbeth⋆	0.611	0.783
Blade⋆	0.601	0.751
Deadpool	0.557	0.679
Batman	0.528	0.695
Little Foot	0.521	0.718
Gandalf	0.486	0.603
Palpatine	0.401	0.562

(b) Best quality score per team

	nDCG@5	
Team	**Qual.**	**Rel.**
Heimdall⋆	0.841	0.639
Skeletor⋆	0.827	0.666
Asterix⋆	0.818	0.663
Elrond⋆	0.817	0.674
Pippin Took⋆	0.814	0.683
Goemon Ishikawa	0.812	0.635
Hua Mulan⋆	0.811	0.620
Dread Pirate Roberts⋆	0.810	0.647
Yeagerists	0.810	0.625
Robin Hood⋆	0.809	0.641
Luke Skywalker	0.808	0.662
Macbeth⋆	0.803	0.608
Athos	0.802	0.637
Jean Pierre Polnareff	0.802	0.633
Swordsman	**0.796**	**0.626**
Shanks⋆	0.795	0.639
Blade⋆	0.763	0.588
Little Foot	0.718	0.521
Batman	0.695	0.528
Deadpool	0.679	0.557
Gandalf	0.603	0.486
Palpatine	0.562	0.401

agreement) are similar to previous studies [21,49,50]. However, we still had a final discussion with all the annotators to clarify potential misinterpretations. Afterwards, each annotator independently judged the results for disjoint subsets of the topics (i.e., each topic was judged by one annotator only).

4.5 Task Results

The results of the runs with the best nDCG@5 scores per participating team are reported in Table 2. Below, we briefly summarize the best configurations of the teams ranked in the top-5 of either the relevance or the quality evaluation. A more comprehensive discussion including all teams' approaches will be part of the forthcoming extended lab overview [9].

Team *Elrond* combined DirichletLM retrieval with a pre-processing pipeline consisting of Krovetz stemming [27], stopword removal using a custom list, removing terms with certain part-of-speech tags, and enriching the document representations using WordNet-based synonyms.

Team *Pippin Took* also used DirichletLM as their basic retrieval model (parameter optimization based on the Touché 2020 judgments) combined with WordNet-based query expansion.

Team *Robin Hood* combined RM3 [28] query expansion with phrase embeddings for retrieval. Their system represents the premise and the conclusion of each argument in two separate vector spaces using the Universal Sentence Encoder [10], and then ranks the arguments based on their cosine similarity to the embedded query.

Team *Asterix* combined BM25 as basic retrieval model with WordNet-based query expansion and a quality-aware re-ranking approach (linear regression model trained on the Webis-ArgQuality-20 dataset [21]). In their system, arguments are ranked based on a combination of the predicted quality score and a normalized BM25 score.

Team *Dread Pirate Roberts* trained a LambdaMART model on the Task 1 relevance labels of Touché 2020 to re-rank the top-100 results of an initial DirichletLM ranking. Using greedy feature selection, they identified the four to nine features with the best nDCG scores in a 5-fold cross-validation setup.

Team *Heimdall* represented arguments using k-means cluster centroids in a vector space constructed using phrase embeddings. Their system combines the cosine similarity of a query to a centroid with DirichletLM retrieval scores, and derives an argument quality score from an SVM regression model that uses $tf \cdot idf$ features and was trained on the overall quality ratings from the Webis-ArgQuality-20 dataset.

Team *Skeletor*, finally, combined a fine-tuned BM25 model with the cosine similarity of passages calculated by a phrase embedding model fine-tuned for question answering. They included pseudo-relevance feedback using the 50 arguments that are most similar in the embedding space to the top-3 initially retrieved arguments. The final retrieval score of a candidate result passage is approximated in their system by its similarity to the relevance feedback passages determined with manifold approximation and summed as the argument's score.

5 Task 2: Argument Retrieval for Comparative Questions

The goal of the Touché 2021 lab's second task was to support individuals making informed decisions in "everyday" or personal comparison situations—in its simplest form for questions such as "Is X or Y better for Z?". Decision making in such situations benefits from finding balanced justifications for choosing one or the other option, for instance, in the form of pro/con arguments.

Table 3. Example topic for Task 2: Argument Retrieval for Comparative Questions.

Number	88
Title	Should I major in philosophy or psychology?
Description	A soon-to-be high-school graduate finds themself 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 similarities, as well as advantages and disadvantages of majoring in either of them (e.g., with respect to career opportunities or gained skills).
Narrative	Relevant documents will overview one of the two majors in terms of career prospects or developed new skills, or they will provide a list of reasons to major in one or the other. Highly relevant documents will compare the two majors side-by-side and help to decide which should be preferred in what context. Not relevant are study program and university advertisements or general descriptions of the disciplines that do not mention benefits, advantages, or pros/cons.

Similar to Task 1, the results of last year's Task 2 participants indicated that improving upon an argument-agnostic BM25 baseline is quite difficult. Promising proposed approaches tried to re-rank based on features capturing "comparativeness" or "argumentativeness".

5.1 Task Definition

Given a comparative question, an approach to Task 2 needed to retrieve documents from the general web crawl ClueWeb12[8] that help to come to an informed decision on the comparison. Ideally, the retrieved documents should be argumentative with convincing arguments for or against one or the other option. To identify arguments in web documents, the participants were not restricted to any system; they could use own technology or any existing argument taggers such as MARGOT [30]. To lower the entry barriers for participants new to argument mining, we offered support for using the neural argument tagger TARGER [13] hosted on our own servers and accessible via an API.[9]

5.2 Data Description

Topics. For the second task edition, we manually selected 50 new comparative questions from the MS MARCO dataset [36] (questions from Bing's search logs) and the Quora dataset [22] (questions asked on the Quora question answering

[8] https://lemurproject.org/clueweb12/.
[9] https://demo.webis.de/targer-api/apidocs/.

website). We ensured to include questions on diverse topics, for example asking about electronics, culinary, house appliances, life choices, etc. Table 3 shows an example topic for Task 2 that consists of a title (i.e., a comparative question), a description of the possible search context and situation, and a narrative describing what makes a retrieved result relevant (meant as a guideline for human assessors). We manually ensured that relevant documents for each topic were actually contained in the ClueWeb12 (i.e., avoiding questions on comparison options not known at the ClueWeb12 crawling time in 2012).

Document Collection. The retrieval corpus was formed by the ClueWeb12 collection that contains 733 million English web pages (27.3 TB uncompressed) crawled by the Language Technologies Institute at Carnegie Mellon University between February and May 2012. For participants of Task 2 who could not index the ClueWeb12 on their side, we provided access to the indexed corpus through the BM25F-based search engine ChatNoir [6] via its API.[10]

5.3 Submitted Approaches

For Task 2, six teams submitted approaches that all used ChatNoir for an initial document retrieval. Most teams then applied a document "preprocessing" on the ChatNoir results (e.g., removing HTML markups) and re-ranked them with feature-based or neural classifiers trained on last year's judgments. Commonly used techniques further included (1) query processing (e.g., lemmatization and POS-tagging), (2) query expansion (e.g., synonyms from WordNet [17], or generated with the word2vec [33] or sense2vec embeddings [47]), and (3) calculating argumentativeness, credibility, or comparativeness scores used as features in the re-ranking. The teams predicted document relevance labels by using a random forest classifier, XGBoost [12], LightGBM [25], or a fine-tuned BERT [14].

5.4 Task Evaluation

Using the CopyCat framework [18], we found that on average 11.6% of the documents in the top-5 results of a run were near-duplicates—a non-negligible redundancy that might have negatively impacted the reliability and validity of an evaluation, since rankings containing multiple relevant duplicates tend to overestimate the actual retrieval effectiveness [19,20]. Following the strategy used in Task 1, we pooled the top-5 documents from the original and the deduplicated runs, resulting in 2,076 unique documents that needed to be judged.

Our eight volunteer annotators (same as for Task 1) labeled a document for its topical relevance (three labels; 0: not relevant, 1: relevant, and 2: highly relevant) and for whether rhetorically well-written arguments [50] were contained (three labels; 0: low quality or no arguments in the document, 1: sufficient quality, and 2: high quality). Similar to Task 1, our eight volunteer assessors went through an initial κ-test on 15 documents from three topics (five documents

[10] https://www.chatnoir.eu/doc/.

Table 4. Results for Task 2 Argument Retrieval for Comparative Questions. The left part (a) shows the evaluation results of a team's best run according to the results' relevance, while the right part (b) shows the best runs according to the results' quality. An asterisk (⋆) indicates that the runs with the best relevance and the best quality differ for a team. The baseline ChatNoir ranking is shown in bold.

(a) Best relevance score per team

	nDCG@5	
Team	**Rel.**	**Qual.**
Katana⋆	0.489	0.675
Thor	0.478	0.680
Rayla⋆	0.473	0.670
Jack Sparrow	0.467	0.664
Mercutio	0.441	0.651
Puss in Boots	**0.422**	**0.636**
Prince Caspian	0.244	0.548

(b) Best quality score per team

	nDCG@5	
Team	**Qual.**	**Rel.**
Rayla⋆	0.688	0.466
Katana⋆	0.684	0.460
Thor	0.680	0.478
Jack Sparrow	0.664	0.467
Mercutio	0.651	0.441
Puss in Boots	**0.636**	**0.422**
Prince Caspian	0.548	0.244

per topic). As in case of Task 1, the observed Fleiss' κ values of 0.46 for relevance (moderate agreement) and of 0.22 for quality (fair agreement) are similar to previous studies [21,49,50]. Again, however, we had a final discussion with all the annotators to clarify some potential misinterpretations. Afterwards, each annotator independently judged the results for disjoint subsets of the topics (i.e., each topic was judged by one annotator only).

5.5 Task Results

The results of the runs with the best nDCG@5 scores per participating team are reported in Table 4. Below, we briefly summarize the best configurations of the teams. A more comprehensive discussion including all teams' approaches will be part of the forthcoming extended lab overview [9].

Team *Katana* re-ranked the top-100 ChatNoir results using an XGBoost [12] approach (overall relevance-wise most effective run) or a LightGBM [25] approach (team Katana's quality-wise best run), respectively. Both approaches were trained on judgments from Touché 2020 employing relevance features (e.g., ChatNoir relevance score) and "comparativness" features (e.g., number of identified comparison objects, aspects, or predicates [11]).

Team *Thor* re-ranked the top-110 ChatNoir results by locally creating an Elasticsearch BM25F index (fields: original and lemmatized document titles, bodies, and argument units (premises and claims) as identified by TARGER; BM25 parameters b and k_1 optimized on the Touché 2020 judgments). This new index was then queried with the topic title expanded by WordNet synonyms [17].

Team *Rayla* re-ranked the top-120 ChatNoir results by linearly combining different scores such as a relevance score, PageRank, SpamRank (all returned by ChatNoir), or an argument support score (ratio of argumentative sentences

(premises and claims) in documents found with their own DistilBERT-based [43] classifier). The weights of the individual scores were optimized in a grid search on the Touché 2020 judgments.

Team *Mercutio* re-ranked the top-100 ChatNoir results returned for the topic titles expanded with synonyms (word2vec [33] or nouns in GPT-2 [42] extensions when prompted with the topic title). The re-ranking was based on the relative ratio of premises and claims in the documents (as identified by TARGER).

Team *Prince Caspian* re-ranked the top-40 ChatNoir results using a logistic regression classifier (features: $tf \cdot idf$-weighted 1- to 4-grams; training on the Touché 2020 judgments) that predicts the probability of a result being relevant (final ranking by descending probability).

6 Summary and Outlook

From the 36 teams that registered for the Touché 2021 lab, 27 actively participated by submitting at least one valid run to one of the two shared tasks:(1) argument retrieval for controversial questions, and (2) argument retrieval for comparative questions. Most of the participating teams used the judgments from the first lab's edition to train feature-based or neural approaches that predict argument quality or that re-rank some initial retrieval result. Overall, many more approaches could improve upon the argumentation-agnostic baselines (DirichletLM or BM25) than in the first year, indicating that progress was achieved. For a potential next iteration of the Touché lab, we currently plan to enrich the tasks by including further argument quality dimensions in the evaluation by focusing on the most relevant/argumentative text passages in the retrieval and by detecting the pro/con stance of the returned results.

Acknowledgments. We are very grateful to the CLEF 2021 organizers and the Touché participants, who allowed this lab to happen. We also want to thank Jan Heinrich Reimer for setting up Doccano, Christopher Akiki for providing the baseline DirichletLM implementation, our volunteer annotators who helped to create the relevance and argument quality assessments, and our reviewers for their valuable feedback on the participants' notebooks.

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: Proceedings of the 42nd German Conference on Artificial Intelligence (KI 2019). pp. 48–59. Springer, Berlin, Heidelberg, New York (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). http://ceur-ws.org/Vol-2696/

3. Kennedy, G.A.: On Rhetoric: A Theory of Civic Discourse. Oxford University Press, Oxford (2006)
4. Bar-Haim, R., Eden, L., Friedman, R., Kantor, Y., Lahav, D., Slonim, N.: From arguments to key points: towards automatic argument summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), pp. 4029–4039. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.371
5. Bar-Haim, R., et al.: 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
6. 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
7. 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). http://ceur-ws.org/Vol-2696/
8. Bondarenko, A., et al.: Overview of Touché 2021: argument retrieval. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12657, pp. 574–582. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_67
9. Bondarenko, A., et al.: Overview of Touché 2021: argument retrieval. In: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, p. (to appear). CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2021)
10. Cer, D., et al.: Universal Sentence Encoder. CoRR **abs/1803.11175** (2018). http://arxiv.org/abs/1803.11175
11. Chekalina, V., Bondarenko, A., Biemann, C., Beloucif, M., Logacheva, V., Panchenko, A.: Which is better for deep learning: python or MATLAB? Answering comparative questions in natural language. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021), pp. 302–311. Association for Computational Linguistics (2021). https://www.aclweb.org/anthology/2021.eacl-demos.36/
12. 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, pp. 785–794. ACM (2016). https://doi.org/10.1145/2939672.2939785
13. 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. Association for Computational Linguistics (2019). https://www.aclweb.org/anthology/P19-3031
14. 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. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
15. Dumani, L., Neumann, P.J., Schenkel, R.: A framework for argument retrieval - ranking argument clusters by frequency and specificity. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12035, pp. 431–445. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_29

16. Dumani, L., Schenkel, R.: Quality aware ranking of arguments. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 335–344. CIKM 2020, Association for Computing Machinery (2020). https://doi.org/10.1007/978-3-030-45439-5_29
17. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
18. Fröbe, M., et al.: 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). ACM (2021). https://doi.org/10.1145/3404835.3463246
19. 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), pp. 1997–2000. ACM (2020). https://doi.org/10.1145/3397271.3401212
20. 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., et al. (eds.) ECIR 2020. LNCS, vol. 12036, pp. 12–19. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_2
21. 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, Online (2020). https://www.aclweb.org/anthology/2020.acl-main.511/
22. Iyer, S., Dandekar, N., Csernai, K.: First Quora Dataset Release: Question Pairs (2017). https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs
23. 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
24. 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
25. 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)
26. 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
27. Krovetz, R.: Viewing morphology as an inference process. In: Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval (SIGIR 1993), pp. 191–202. ACM (1993). https://doi.org/10.1145/160688.160718
28. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2001), pp. 120–127. ACM (2001). https://doi.org/10.1145/383952.383972

29. Levy, R., Bogin, B., Gretz, S., Aharonov, R., Slonim, N.: Towards an argumentative content search engine using weak supervision. In: 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/
30. Lippi, M., Torroni, P.: MARGOT: a web server for argumentation mining. Expert Syst. Appl. **65**, 292–303 (2016). https://doi.org/10.1016/j.eswa.2016.08.050
31. 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/
32. Mass, Y., et al.: 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
33. 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
34. 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
35. Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., Liang, X.: doccano: Text Annotation Tool for Human (2018). https://github.com/doccano/doccano
36. Nguyen, T., et al.: MS MARCO: a human generated machine reading comprehension dataset. In: Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016). CEUR Workshop Proceedings, vol. 1773. CEUR-WS.org (2016). http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf
37. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999–66, Stanford InfoLab (1999). http://ilpubs.stanford.edu:8090/422/
38. 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
39. 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
40. 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
41. 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

42. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
43. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. CoRR abs/1910.01108 (2019). http://arxiv.org/abs/1910.01108
44. 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
45. 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-HLT 2018), pp. 21–25. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/n18-5005
46. 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
47. 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
48. Wachsmuth, H., et al.: 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
49. Wachsmuth, H., et al.: 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)
50. 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
51. Wachsmuth, H., et al.: 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
52. 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
53. 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/