Webis at TREC 2020: Health Misinformation Track

Extended Abstract

Janek Bevendorff [1,*] Alexander Bondarenko [2,*] Maik Fröbe [2,*] Sebastian Günther [2,*]
Michael Völske [1,*] Benno Stein [1] Matthias Hagen [2]

[1] Bauhaus-Universität Weimar
⟨first⟩.⟨last⟩@uni-weimar.de

[2] Martin-Luther-Universität Halle-Wittenberg
⟨first⟩.⟨last⟩@informatik.uni-halle.de

ABSTRACT

This paper gives a brief overview of the Webis group's participation in the TREC 2020 Health Misinformation track. We use the search engine ChatNoir for baseline retrieval in our two approaches: (1) axiomatically re-ranking the top-20 initial results for argumentative topics/queries, and (2) formulating keyqueries to retrieve relevant documents at the top ranks. Our axiomatic re-ranking uses three axioms that capture argumentativeness for the re-ranking, while for the keyqueries approach, we used low-effort manual pilot judgments to identify a few relevant documents per topic.

1 INTRODUCTION

Search results containing wrong, unreliable, or misleading information can be harmful to a searcher, who takes the information for granted—especially in scenarios like health search. The Health Misinformation track focuses on information needs like "Can ibuprofen worsen COVID-19?", upon whose results searchers might make wrong decisions that negatively affect their health. Our approaches to the Health Misinformation track scenario try to rank documents higher that provide correct information from credible sources.

Our runs use the Elasticsearch-based search engine ChatNoir [2], which indexes the Common Crawl News corpus by pre-processing the raw WARC files to extract the main content, the language, and metadata such as keywords, headings, hostnames, etc. We submitted the following seven runs: (1–2) two runs using either the topic's title or its description as the ChatNoir query, (3) one based on axiomatic re-ranking [3, 4], and (4–7) four manual runs based on keyqueries computed for manual pilot judgments.

2 MANUAL PILOT JUDGMENTS

In a manual pilot judgment phase (budget: six minutes per topic) we tried to obtain for each topic a small number of "target" documents with correct answers. Our motivation is to use the identified target documents as relevance feedback for our two keyquery-based runs.

Four annotators did our pilot judgments using the same instructions; three annotators got 10 non-overlapping topics each, one got the remaining 20 topics. The annotators' task was to identify at least two documents that likely provide a correct answer to the given information need. The annotators started by carefully reading the full topic (including all provided fields: title, description, answer, evidence, and narrative) to understand which documents would provide correct answers to the described information need. Then, the annotators used the web interface of ChatNoir to identify documents with correct answers for the topic. They were allowed to formulate any query they want but should only look at the ChatNoir result pages but not click on any link (i.e., only result URLs, titles, and snippets were shown). The snippets contained up to 300 characters with Elasticsearch's query term highlighting. Given the tight timing constraints, we did not ask our annotators to assess a documents' credibility but just whether the content will likely be relevant given the topic description. Although the annotators could stop early after identifying two target documents, we collected 3 target documents per topic on average (with a maximum of 11 target documents for Topic 41: `Hib vaccine COVID-19`).

Table 1: Assessment of our pilot judgments used for our keyquery-based runs (Pilot) compared to all runs submitted to the TREC Health Misinformation track (All).

	Useful (%)		Answer (%)			Credible (%)	
	Yes	No	Correct	Not Correct	No Answer	Yes	No
Pilot	83.9	16.1	62.7	5.2	32.1	84.3	15.7
All	34.0	66.0	40.6	11.1	48.3	81.7	18.3

Table 1 provides an overview of the official NIST assessors' judgments for the target documents from our pilot judgments and all judgments for runs submitted to the TREC 2020 Health Misinformation track. Our pilot judgments' quality is pretty good: 84% of our target documents are judged as useful—despite the minimal effort of invested time. However, only 63% of our target documents provide the correct answer (compared to 41% correct answers in the results of all submitted runs). A critical observation is that our pilot judgments' "accuracy" advantage compared to all runs' results decreases from simplistic document properties like "usefulness" (50 percentage points advantage), over "correct answers" (22 percentage points advantage), to credibility (3 percentage points advantage only). This indicates that our pilot judgments could not really support assessing a result's credibility since probably a document's URL, title, and short snippet snippet are not sufficient but rather the whole content is needed (e.g., sources for claims given in a document, etc.).

3 WEBIS RUNS

We submit seven runs that can be divided into the three groups: (1) baseline retrieval with ChatNoir, (2) axiomatic re-ranking with

[*] These five authors contributed to the paper equally.

TREC 2020, November 16–20, 2020, Gaithersburg, Md.
2020. Webis network [webis.de].

argumentative axioms, and (3) manual runs based on pilot judgments. All runs use the BM25F-based search engine ChatNoir [2].

3.1 Baseline Retrieval With ChatNoir

We used ChatNoir [2] as the basis for all our runs. ChatNoir leverages a large-scale Elasticsearch cluster with 130 nodes to offer a freely accessible search interface for the two ClueWebs and two Common Crawl snapshots,[1] together about 5 billion web pages.

We used ChatNoir's indexing pipeline to index the Common Crawl News corpora's documents by processing the raw WARC files using main content extraction, language detection, and metadata extraction (keywords, headings, hostnames, etc.). During retrieval, we used ChatNoirs existing weighting scheme for the two Common Crawl snapshots, which combines BM25 scores of multiple fields (title, URL, keywords, main content, and the full document). ChatNoir comes with a web interface and a REST-API, and we have used the REST-API for all our runs.

In the end, we submitted two standalone ChatNoir runs, and five runs that use ChatNoir as baseline retrieval systems. The first standalone ChatNoir run uses the given title as a query, and the second run uses the description as the query against ChatNoir.

3.2 Argumentative Axiomatic Re-ranking

We create and submit one run based on re-ranking the top-20 initial retrieved results with ChatNoir using argumentative axioms. Largely, we apply the same re-ranking strategy used in our previous TREC participation [3, 4].

3.2.1 Identifying Argumentative Queries. Based on the assumption that users issuing argumentative queries might prefer documents containing argumentation directly [3, 4], we first identify which Health Misinformation track's topics are argumentative. For that, we manually inspected all topics. Given their medical COVID-19-related nature, we concluded that all of them could be labeled as argumentative queries (i.e., relevant results containing some form of argumentation might be perceived as more relevant/helpful).

3.2.2 Re-ranking Axioms. To re-rank the top-20 BM25F-based retrieval results we use the three axioms which favor more argumentative documents from our previous years' Common Core and Decision tracks contributions [3, 4].

Retrieval axioms (i.e., formally defined constraints applied to retrieval models) have been developed within the axiomatic thinking in information retrieval [1] to define some algorithmic heuristics that good retrieval models should follow. Traditionally, such constraints were developed to account the relevance of retrieved documents to the respective queries. The basic example is the term-frequency axiom TFC1 [6], which states that given two documents of the same length and a single-term query, the document with more occurrences of the query term should receive a higher ranking score from some query-document scoring function. We follow the ideas of axiomatic thinking and address the argumentative nature of queries and documents by using the three axioms (used also in our previous TREC participation) that capture document argumentativeness. Note that we relax the precondition of documents' lengths equality to ensure the axioms' applicability to the real web documents. We formally define our axioms as follows:

Axiom ArgUC (Argumentative Units Count). The general idea of the ArgUC axiom is to favor documents that contain a larger number of argumentative units.

Formalization. Let Q be an argumentative query, D_1 and D_2 be two retrieved documents with $count_{\text{Arg}}(D_1)$ and $count_{\text{Arg}}(D_2)$ argumentative units counts in documents, and let $\approx_{10\%}$ indicate "equality" up to a 10% difference. If $length(D_1) \approx_{10\%} length(D_2)$ and $count_{\text{Arg}}(D_1) > count_{\text{Arg}}(D_2)$, then $rank(D_1, Q) > rank(D_2, Q)$.

Axiom QTArg (Query Term Occurrence in Argumentative Units). Retrieved documents usually consist of argumentative and non-argumentative units or text passages. The general idea of the QTArg axiom is to favor documents where the query terms appear closer to argumentative units.

Formalization. Let $Q = \{q\}$ be an argumentative single-term query, D_1 and D_2 be two retrieved documents, and let Arg_D be the set of argumentative units of a document D. If $length(D_1) \approx_{10\%} length(D_2)$ and $q \in A_{D_1}$ for some $A_{D_1} \in Arg_{D_1}$ but $q \notin A_{D_2}$ for all $A_{D_2} \in Arg_{D_2}$, then $rank(D_1, Q) > rank(D_2, Q)$.

Axiom QTPArg (Query Term Position in Argumentative Units). Following the general observation that in relevant documents the query terms occur closer to the beginning [10, 12], the QTPArg axiom will favor documents where the first appearance of a query term in an argumentative unit is closer to the beginning of the document.

Formalization. Let $Q = \{q\}$ be an argumentative single-term query, D_1 and D_2 be two retrieved documents, and let the first position in an argumentative unit of a document D where the term q appears be denoted by $1^{st}position(q, Arg_D)$. If $length(D_1) \approx_{10\%} length(D_2)$ and $1^{st}position(q, Arg_{D_1}) < 1^{st}position(q, Arg_{D_2})$, then $rank(D_1, Q) > rank(D_2, Q)$.

3.2.3 Argumentative Unit Detection. The three argumentative axioms are based on argumentative units in documents. To detect argumentative units, we use the BiLSTM-CNN-CRF argument tagging tool TARGER [5] that is available via the API.[2] TARGER takes as input a raw text and returns a text tagged with markers indicating the beginning and the end of argument premises and claims (argumentative units).

3.2.4 Final Run. In addition to the three argumentative axioms, we also employ an axiom ORIG [9] that simply returns the preferences of the baseline retrieval system's ranking—BM25F. We do this to balance between document argumentativeness and relevance to avoid more argumentative but less relevant documents being ranked higher. The four different axioms (including ORIG) are weighted to linearly combine the respective preference matrices following the original axiomatic re-ranking pipeline [9]. The weight of an axiom then directly influences its impact on the document re-ranking. For simplicity, we apply the *argumentative re-ranking* vs. *original ranking* axiom weighting strategy, such that document positions are swapped in the ranking iff all three argumentative axioms agree to overrule the ORIG preference.

[1] We have indexed a 2015 and 2017 snapshot, available at: https://commoncrawl.org

[2] https://demo.webis.de/targer-api/apidocs/

3.3 Manual Runs

We used the target documents identified in the pilot judgments to produce four manual runs. The first two manual runs use the two standalone ChatNoir runs using the topics' title (respectively the topics' description) as the query against ChatNoir, and artificially move the target documents from the pilot judgments to the top of the ranking. This ensures that we obtain complete judgments for our pilot judgments. The other two manual runs use the pilot judgments as relevance feedback to calculate keyqueries [7].

Our target documents are intended to be useful and answer the information need. Hence, our underlying working hypothesis is that the keyqueries retrieve documents similar to the target documents (i.e., useful with the correct answer) at high positions. This hypothesis is motivated by applications of keyqueries in scholarly search, where they can retrieve related work effectifely [8].

A keyquery [7] retrieves at least l documents with m documents from a given set of target documents within the top k results. The parameter l controls the level of the keyqueries' generality, and we set $l = 25$ to ensure that each query retrieves new documents and does not overfit by retrieving exactly the target documents. We further set $m = 1$ and $k = 20$ to ensure that we obtain many keyqueries. I.e., a query is a keyquery when it retrieves more than 25 documents, and at minimum one target document from our pilot judgments is within the top 20 documents.

We create candidate queries by leveraging Elasticsearch's term vectors API. We combine two strategies to create query candidates. First, we use each target document to produce 32 candidates per target document by selecting the power set of the five terms from the document with the highest BM25 scores on the main-content field. Additionally, we select 256 candidates per set of target documents by selecting the power set of the eight terms with the highest BM25 scores combined over all target documents. We verify all candidate queries and remove candidates that are no keyqueries for our parameters (with m=1, k=20, l=25) for the target documents.

We use a greedy algorithm to combine the identified keyqueries to produce the runs. Starting with the set of target documents from our pilot judgments for a topic, we select the keyquery with the highest nDCG (we consider all target documents as relevant) and remove retrieved documents from the set of target documents. Until the set of target documents is empty, we iteratively select the query with the highest nDCG on the remaining target documents. We combine the selected keyqueries using the topic's title (respectively, the description) as an additional mandatory query with team-draft-interleaving [11], which produces our two keyquery runs.

4 CONCLUSION

We implemented two approaches—one with axiomatic re-ranking and one leveraging keyqueries—to support the users' decision-making process in medical contexts. Our two approaches aim to move useful and credible documents with correct answers to the top of the ranking. We conducted some pilot judgments that confirm the importance of the underlying task. We could identify relevant documents from the users' perspective but could not distinguish credible/incredible documents on the search engine result pages.

We believe that an important line of future work for both of our implemented approaches is to leverage (a subset of) the official judgments to tune the approaches' parameters. We selected the parameters only using our experience, which may be entirely wrong for retrieval tasks in the quickly changing COVID-19 domain.

ACKNOWLEDGMENTS

This work has been partially supported by the DFG through the project "ACQuA: Answering Comparative Questions with Arguments" (grant HA 5851/2-1) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999).

REFERENCES

[1] Enrique Amigó, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. Axiomatic Thinking for Information Retrieval: And Related Tasks. In *Proceedings of the 40th International ACM SIGIR 2017 Conference on Research and Development in Information Retrieval*. 1419–1420.

[2] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018) (Lecture Notes in Computer Science)*, Leif Azzopardi, Allan Hanbury, Gabriella Pasi, and Benjamin Piwowarski (Eds.). Springer, Berlin Heidelberg New York.

[3] Alexander Bondarenko, Maik Fröbe, Vaibhav Kasturia, Michael Völske, Benno Stein, and Matthias Hagen. 2019. Webis at TREC 2019: Decision Track. In *28th International Text Retrieval Conference (TREC 2019) (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST), 4.

[4] Alexander Bondarenko, Michael Völske, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. 2018. Webis at TREC 2018: Common Core Track. In *27th International Text Retrieval Conference (TREC 2018) (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST), 3.

[5] Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural Argument Mining at Your Fingertips. In *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Martha R. Costa-jussà and Enrique Alfonseca (Eds.). Association for Computational Linguistics, 195–200. https://www.aclweb.org/anthology/P19-3031

[6] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*. 49–56. DOI:http://dx.doi.org/10.1145/1008992.1009004

[7] Tim Gollub, Matthias Hagen, Maximilian Michel, and Benno Stein. 2013. From Keywords to Keyqueries: Content Descriptors for the Web. In *36th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2013)*, Cathal Gurrin, Gareth J.F. Jones, Diane Kelly, Udo Kruschwitz, Maarten de Rijke, Tetsuya Sakai, and Páraic Sheridan (Eds.). ACM, 981–984. DOI:http://dx.doi.org/10.1145/2484028.2484181

[8] Matthias Hagen, Anna Beyer, Tim Gollub, Kristof Komlossy, and Benno Stein. 2016. Supporting Scholarly Search with Keyqueries. In *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016) (Lecture Notes in Computer Science)*, Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello (Eds.), Vol. 9626. Springer, Berlin Heidelberg New York, 507–520. DOI:http://dx.doi.org/10.1007/978-3-319-30671-1_37

[9] Matthias Hagen, Michael Völske, Steve Göring, and Benno Stein. 2016. Axiomatic Result Re-Ranking. In *25th ACM International Conference on Information and Knowledge Management (CIKM 2016)*. ACM, 721–730.

[10] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to Match Using Local and Distributed Representations of Text for Web Search. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017*. 1291–1299.

[11] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How does click-through data reflect retrieval quality?. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*. ACM, Napa Vallay, California, USA, 43–52. DOI:http://dx.doi.org/10.1145/1458082.1458092

[12] Adam D. Troy and Guo-Qiang Zhang. Enhancing Relevance Scoring with Chronological Term Rank. In *Proceedings of the 30th Annual International ACM SIGIR 2007 Conference on Research and Development in Information Retrieval*. 599–606.