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

We give a brief overview of the Webis group's participation in the TREC 2020 Health Misinformation track. The baseline retrieval results of our search engine ChatNoir (BM25F-based) are re-ranked in two different 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, while for the keyqueries approach, we use low-effort manual pilot judgments to identify several relevant documents per topic.

1 INTRODUCTION

Search results that contain wrong, unreliable, or misleading information can be harmful to searchers who take the information for granted—especially in scenarios of health search. The Health Misinformation track at TREC 2020 focuses on information needs like "Can ibuprofen worsen COVID-19?", for which the results might lead searchers to wrong decisions that negatively affect their health. Our approaches to the Health Misinformation track scenario attempt to rank documents higher that provide correct and justified information from credible sources.

We submitted ranked result lists using the Elasticsearch-based search engine ChatNoir [2], which indexes the Common Crawl News corpus using BM25F [13] as the retrieval model and preprocessing the raw WARC files to extract the main content and the metadata such as keywords, headings, host names, etc. Based on this baseline retrieval system, we submitted the following seven runs: (1–2) two runs using either the topic's title or its description as the query to ChatNoir, (3) one run based on axiomatic re-ranking [3, 4], and (4–7) four manual runs based on manual pilot judgments that are used to automatically formulate keyqueries.

2 MANUAL PILOT JUDGMENTS

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

Four annotators did the pilot judgments using the same annotation instructions. Three of them got 10 non-overlapping topics each and one got the remaining 20 topics. The annotators' task was to identify at least two documents which most likely provide a correct answer to the question / topic title. The annotators started by carefully reading the full topic: 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 potentially correct answers. They were allowed to formulate their own queries (i.e., re-formulating the topic titles that were given as questions) and were instructed to look only at the ChatNoir result page and to not click any links (i.e., only result URLs, titles, and snippets were shown). The snippets contained up to 300 characters with query term highlighting by Elasticsearch's functionality. Given the restricting timing constraints (six minutes per topic), we did not ask our annotators to assess a documents' credibility but only whether the content is likely to be correct given the topic description. Although the annotators were allowed to stop as early as they had identified 2 target documents, we collected 3 target documents per topic on average (maximum of 11 target documents for Topic 41: `Hib vaccine COVID-19`).

Table 1: Assessment of our pilot judgments used for the 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 Health Misinformation track judgments for the target documents from our pilot judgments compared to all Health Misinformation track judgments for all runs submitted by any participant to the TREC 2020 Health Misinformation track. Our pilot judgments show a pretty good quality even though only the URLs, titles, and snippets were inspected by our annotators given some limited time budget: 84% of our target documents are judged as useful. 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 the pilot judgments' "accuracy" advantage over all runs diminishes with the complexity of the quality criteria. Although the pilot judgments achieve a much higher "usefulness" score (50 percentage points advantage), we the advantage decreases for "correct answers" (22 points advantage), and almost diminishes for "credible" documents (3 points advantage only). This indicates

*These five authors contributed to the paper equally.

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

that our pilot judgments could not support the assessment of a result's credibility: a document's URL, title, and short snippet, that our annotators were allowed to use, are simply not sufficient but the whole content is needed for a proper assessment (e.g., sources of 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 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 ClueWeb and two Common Crawl snapshots, about 5 billion web pages altogether.[1]

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

We submitted two standalone ChatNoir runs and five runs that use ChatNoir for initial retrieval before re-ranking. The first ChatNoir run uses only the title as a query and the second run uses the description as a query.

3.2 Argumentative Axiomatic Re-ranking

We created and submitted one run based on a re-ranking of the top-20 initially retrieved ChatNoir results using argumentative axioms. We largely 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, i.e., queries that demand justification of the retrieved information, might prefer documents containing argumentation [3, 4], we first identify which of the Health Misinformation track's topics are argumentative. For this, 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 ChatNoir retrieval results, we use the three axioms which re-rank argumentative documents higher from our previous years' Common Core and Decision track 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 which good retrieval models should fulfill. 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 [7], 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 any 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 that capture document argumentativeness (and which were also used in our previous TREC submissions). Note that we relax the precondition of document length equality to ensure the axioms' applicability to 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 [11, 14], the QTPArg axiom favors 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 an API.[2] TARGER takes as input a raw text and returns markers indicating the beginning and the end of argument premises and claims (argumentative units).

3.2.4 Actual Run. In addition to the three argumentative axioms, we also employ the axiom ORIG [10], which simply returns the preferences of the baseline retrieval system's ranking—BM25(F). 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)

[1] We have indexed the 2015 and 2017 snapshots available at: https://commoncrawl.org

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

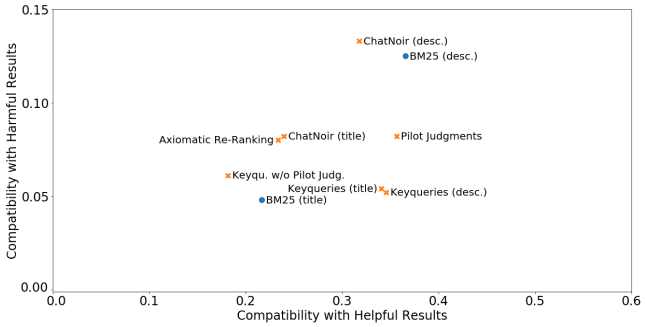

Figure 1: Compatibility of our runs (crosses) and the official BM25 baselines (circles) with the help/harm preferences [6].

are weighted to linearly combine the respective preference matrices following the original axiomatic re-ranking pipeline [10]. 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 in the initial ranking are swapped iff all three argumentative axioms agree to overrule the ORIG preference.

3.3 Manual and Keyquery Runs

Four manual runs employ the target documents identified in the pilot judgments. The first manual run simply moves the target judgments identified in the pilot judgments to the top of ChatNoir's ranking for the title query. The other three manual runs use the pilot judgments as target documents to formulate keyqueries [8]. Slightly extending the original definition [8], we view a query as a keyquery for a given target document set iff it retrieves at least m of the target documents in the top k of at least l results. The parameter l controls the level of the keyqueries' generality, the parameter k its fit to the target documents. We set $k = 20$, $l = 25$, and $m = 1$ to ensure that we find some keyqueries that each retrieve new documents and do not overfit by retrieving exactly the target documents only.

Our target documents labeled manually are intended to be useful and to correctly answer the question formulated in a topic. Our hypothesis is that keyqueries retrieve documents similar to the target documents at high ranks. This hypothesis is motivated by the usage of keyqueries in scholarly search, where they are effective for retrieving related work [9].

To create candidate keyqueries, we leverage Elasticsearch's term vector API using two strategies. In the first approach, for each individual target document for some topic, we extract the five terms with highest BM25 scores on the main content field. With these 5 terms, we formulate all 32 possible combinations ($2^5 = 32$) as keyquery candidates per target document. In the second approach, we select from the combination of the topic's target documents the eight terms with the highest BM25 scores and formulate all 256 possible combinations from these terms as candidate keyqueries. In both approaches, we verify all candidate queries and remove candidates that are not keyqueries for the target documents (parameters set to $k = 20$, $l = 25$, and $m = 1$).

We use a greedy algorithm for combining the identified keyqueries to produce the runs. Starting with a topic's set of target documents from our pilot judgments, we select the keyquery with the highest nDCG (just considering the target documents as relevant) and remove the retrieved documents from the set of target documents. We iteratively select the query with the highest nDCG on the remaining target documents until the set of target documents is empty. We combine the selected keyqueries using the topic's title (or, respectively, the description) as additional query with team-draft-interleaving [12], which produces two keyquery runs.

Additionally, we also submit a run that contains a mixture of lower-ranked documents retrieved by our keyqueries. From these lower-ranked results, we removed all target documents since this run should simply ensure that sufficiently many new documents retrieved by our keyqueries are judged.

Table 2: Overview of the compatibility and nDCG scores of our runs and the official BM25 baselines (scores reported by the Health Misinformation track, approaches sorted by help - harm [6]).

Approach	Compatibility		nDCG on Binary Qrels			
	Help	Harm	Useful	Correct	Credible	All
Keyqueries (desc.)	0.334	0.052	0.258	0.227	0.263	0.212
Keyqueries (title)	0.331	0.054	0.341	0.280	0.332	0.264
Pilot Judgments	0.357	0.082	0.443	0.337	0.420	0.318
BM25 (desc.)	**0.366**	0.125	**0.605**	**0.495**	**0.574**	**0.483**
ChatNoir (desc.)	0.318	0.133	0.385	0.316	0.360	0.305
BM25 (title)	0.217	**0.048**	0.461	0.327	0.441	0.318
ChatNoir (title)	0.240	0.082	0.413	0.290	0.383	0.275
Ax. Re-ranking	0.234	0.080	0.404	0.280	0.376	0.265
Keyqu. w/o Pilot Judg.	0.182	0.061	0.198	0.150	0.188	0.136

4 EVALUATION

In the Health Misinformation track, the retrieval effectiveness of the submitted rankings is evaluated with novel compatibility measures and multiple variants of nDCG [6]. To this end, NIST assessors labeled the usefulness, correctness, and credibility of documents. The compatibility measure combines these three aspects into preference orderings for the document helpfulness and harmfulness.

Figure 1 shows our submitted runs and the official BM25 baselines in a "compatibility" plot. With only the title as the query, our ChatNoir baseline systems produces more harmful rankings than the BM25 baseline while for the descriptions as queries, the ChatNoir rankings are slightly more harmful and also less helpful than the official baseline run. Axiomatically re-ranking the ChatNoir results with the title as query has minimal effects on the compatibility (helpfulness and harmfulness slightly decrease). But moving the target documents from the pilot judgments to the top of ChatNoir's title ranking substantially increases the helpfulness, while formulating keyqueries for the pilot judgments increases helpfulness and reduces harmfulness at the same time.

Table 2 shows our runs and the official baselines ordered by compatibility along with the nDCG scores for all aspects judged by the NIST assessors. With a compatibility focus, our keyquery run with pilot judgments produces a more helpful and less harmful ranking than the official BM25 baselines. But evaluating the ranking effectiveness with nDCG shows a different picture: our runs have

worse nDCG scores than the official BM25 baselines for all aspects and the combination of the aspects.

5 CONCLUSION

To participate in the TREC 2020 Health Misinformation track, we submitted runs falling into three types: ChatNoir-based baselines using the topic title or the topic description as queries, and several re-ranking approaches with argumentative axioms and with keyqueries based on manual pilot judgments. Our re-ranking approaches aim to move useful and credible documents with correct answers to the top of the initial ChatNoir BM25F rankings. The results show that the pilot annotation purely based on documents' URLs, titles, and snippets is challenging. Our pilot annotators were able to identify relevant documents from the searcher perspective, but could not reliably distinguish between credible and non-credible documents. Still, the usage of pilot judgments to guide the keyquery formulation helped to improve the results' helpfulness while at the same time decreasing their harmfulness.

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] Charles L. A. Clarke, Maria Maistro, and Mark D. Smucker. 2020. Overview of the TREC 2020 Health Misinformation Track. In *Proceedings of The Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Conference, November 16-20, 2020*. https://trec.nist.gov/act_part/conference/papers/OVERVIEW.H.pdf

[7] 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

[8] 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

[9] 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

[10] 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.

[11] 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.

[12] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How does clickthrough 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

[13] Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM '04)*. ACM, New York, NY, USA, 42–49. DOI: http://dx.doi.org/10.1145/1031171.1031181

[14] 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.