Axiomatic Re-Ranking for Argument Retrieval

Alexander Bondarenko, Maik Fröbe, Jan Heinrich Reimer, and Matthias Hagen

Martin-Luther-Universität Halle-Wittenberg
<first>.<last>@informatik.uni-halle.de

Abstract In this extended abstract, we describe our ongoing research of applying weighted combined axiomatic preferences to improve the search result rankings for argumentative queries, i.e., queries for which the results should include good argumentation.

1 Introduction

The general task of information retrieval is to rank documents that help a user answer their information need expressed as a query. In case of argumentative queries (i.e., explicitly or implicitly asking for argumentation), the ranking must not only consider the topical relevance of a document but also the logical cogency, convincingness, and rhetorical quality of the contained arguments [16, 1, 2].

A line of research in information retrieval that theoretically analyzes constraints for good rankings is the axiomatic approach. Over the last two decades over 20 basic constraints were formulated as so-called "axioms". For example, the axiom TFC1 states that, from two documents of the same length, a good ranking must rank higher the document with more query term occurrences [5]. Besides theoretically analyzing several retrieval models, the suggested axioms have recently also been used in more practical scenarios: (1) re-ranking an initial baseline ranking according to combined and weighted axiom preferences [6], (2) using axiom decisions as regularization loss in neural ranking models [13], and (3) explaining neural ranker preferences by axiom combinations [11, 3, 15].

2 Retrieval Axioms for Argumentativeness

Applying the axiomatic re-ranking approach [6], our idea for argumentative queries is to re-rank argument-agnostic baseline retrieval results according to the preferences of axioms that focus on argumentativeness and writing style. In our current prototype, we have implemented the following four such axioms.

Axiom ArgUC (Argumentative Units Count)
Idea: Favor documents with more argumentative units.
Formalization: Let q be an argumentative query, d_1 and d_2 be two retrieved documents, and let $count_{\text{arg}}(d)$ be the number of argumentative units in document d. If $length(d_1) = length(d_2)$ and $count_{\text{arg}}(d_1) > count_{\text{arg}}(d_2)$, then $d_1 >_{\text{ArgUC}} d_2$.

Axiom QTArg (Query Term Occurrence in Argumentative Units)
Idea: Favor documents with query terms in / closer to argumentative units.
Formalization: Let $q = \{t\}$ be an argumentative single-term query, d_1 and d_2 be two retrieved documents, and let A_d be the set of argumentative units of document d. If $length(d_1) = length(d_2)$ and $t \in a$ for some $a \in A_{d_1}$ but $t \notin a'$ for all $a' \in A_{d_2}$, then $d_1 >_{\text{QTArg}} d_2$.

Axiom QTPArg (Query Term Position in Argumentative Units)
Idea: Favor documents where the first occurrence of a query term in an argumentative unit is closer to the beginning of the document. (General observation for retrieval: query terms occur "earlier" in relevant documents [14, 9].)
Formalization: Let $q = \{t\}$ be an argumentative single-term query, d_1 and d_2 be two retrieved documents, and let $1^{st}pos(t, d)$ be the first position of term t in an argumentative unit of document d. If $length(d_1) = length(d_2)$ and if $1^{st}pos(t, d_1) < 1^{st}pos(t, d_2)$, then $d_1 >_{\text{QTPArg}} d_2$.

Axiom aSL (Average Sentence Length)
Idea: Favor documents with avg. sentence length between 12 and 20 words. (General observation for text readability / good writing style [8, 10].)
Formalization: Let q be an argumentative query, d_1 and d_2 be two retrieved documents, and let $sentLength(d)$ be the average sentence length (in words) of document d. If $length(d_1) = length(d_2)$, $12 \leq sentLength(d_1) \leq 20$, and $sentLength(d_2) < 12$ or $sentLength(d_2) > 20$, then $d_1 >_{\text{aSL}} d_2$.

As also previous studies suggested [6, 15, 3], we relax the axioms' document length preconditions in the actual re-ranking pipeline to consider documents with a length difference of at most 10% as of the same length, and extend the single-term query axioms to also cover multi-term queries. To identify argumentative units in documents (i.e., claims and premises), we use the `TARGER` toolkit [4].[1]

3 Evaluation

We have evaluated our argumentative re-ranking pipeline in the scenarios of two TREC tracks:[2] the TREC 2018 Common Core track (Washington Post corpus; 728,626 news articles) and the TREC 2019 Decision track (ClueWeb12-B13 corpus; 52,343,021 English web pages). From the 50 topics for each of these tracks, we have manually selected the ones that could be interpreted as argumentative queries in the sense that results with good argumentation could be more helpful.

As the initial argumentation-agnostic retrieval model, we use BM25F [12], a variant of BM25 that can take multiple fields into account such that query term matches in the title can be weighted higher than in the body. In addition to the four argumentativeness axioms, we also employ an axiom ORIG [6] that simply returns the preferences corresponding to the baseline retrieval system's ranking. For every pair of documents in the top-50 results of the BM25F baseline,

[1] https://demo.webis.de/targer/
[2] https://trec.nist.gov/

a weighted linear combination of the axiom preferences then is used to decide whether the documents' order should be swapped (note that some document pairs might not yield an axiom preference). We manually apply the following three weighting schemes.

Equal Weights: All axioms get the same weight. This way, any agreement of the preferences of a pair of the new axioms may overrule the ORIG axiom preference when no other axiom "supports" the ORIG preference.

Majority Voting: The axioms are assigned weights such that document pairs are re-ranked iff the majority of the new axioms (at least 3 out of 4 axioms) agree to overrule the ORIG preference.

Total Agreement: The axioms are assigned weights such that document pairs are re-ranked only when all the new axioms agree. It is not necessary for all axioms to have the same weight, although all of them have to be in agreement to overrule the ORIG axiom.

Our evaluations of our three axiom weighting schemes on the two TREC tracks show that different weighting schemes indeed yield different rankings with varying effectiveness. For some queries, applying axiomatic re-ranking improves the retrieval effectiveness compared to the BM25F baseline by up to 20% (e.g., `airport security`, or `euro opposition`). For some other queries, the effectiveness drops of up to 20% for queries like `amazon rain forest` or `marijuana potency`. By further manually varying the axiom weights, we have observed different effectiveness results such that learning optimal axiom weights from a set of labeled training data is an interesting direction for future work. To ensure an efficient training process, we are currently implementing an axiomatic re-ranking module in the Capreolus information retrieval toolkit [17].

4 Conclusion and Outlook

Our initial results show that an axiomatic re-ranking can improve the result quality for argumentative queries. This motivates further research in three directions: (1) automatically identifying argumentative queries, i.e., deciding whether argumentative re-ranking is promising for a given query; (2) automatically assigning the axiom weights; and (3) developing and adding new retrieval axioms that capture more fine-grained aspects of argumentativeness and argument quality.

Acknowledgments

This work was 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).

Bibliography

[1] Bondarenko, A., Fröbe, M., Beloucif, M., Gienapp, L., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2020: Argument Retrieval. In: Proceedings of the Working Notes of CLEF 2020. CEUR, vol. 2696.

[2] Bondarenko, A., Gienapp, L., Fröbe, M., Beloucif, M., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2021: Argument Retrieval. In: Proceedings of the Working Notes of CLEF 2021. CEUR.

[3] Câmara, A., Hauff, C.: Diagnosing BERT with Retrieval Heuristics. In: Proceedings of ECIR 2020. pp. 605–618. Springer.

[4] Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., Panchenko, A.: TARGER: Neural Argument Mining at Your Fingertips. In: Proceedings of ACL 2019. pp. 195–200. ACL.

[5] Fang, H., Tao, T., Zhai, C.: A Formal Study of Information Retrieval Heuristics. In: Proceedings of SIGIR 2004. pp. 49–56. ACM.

[6] Hagen, M., Völske, M., Göring, S., Stein, B.: Axiomatic Result Re-Ranking. In: Proceedings of CIKM 2016. pp. 721–730. ACM.

[7] Järvelin, K., Kekäläinen, J.: Cumulated Gain-based Evaluation of IR Techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002).

[8] Markel, M.: Technical Communication. 9th ed. Bedford/St Martin's (2010)

[9] Mitra, B., Diaz, F., Craswell, N.: Learning to Match Using Local and Distributed Representations of Text for Web Search. In: Proceedings of WWW 2017. pp. 1291–1299. ACM.

[10] Newell, C.: Editing Tip: Sentence Length (2014)

[11] Rennings, D., Moraes, F., Hauff, C.: An Axiomatic Approach to Diagnosing Neural IR Models. In: Proceedings of ECIR 2019. pp. 489–503. Springer.

[12] Robertson, S.E., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (2009).

[13] Rosset, C., Mitra, B., Xiong, C., Craswell, N., Song, X., Tiwary, S.: An Axiomatic Approach to Regularizing Neural Ranking Models. In: Proceedings of SIGIR 2019. pp. 981–984. ACM.

[14] Troy, A.D., Zhang, G.: Enhancing Relevance Scoring with Chronological Term Rank. In: Proceedings of SIGIR 2007. pp. 599–606. ACM.

[15] Völske, M., Bondarenko, A., Fröbe, M., Stein, B., Singh, J., Hagen, M., Anand, A.: Towards Axiomatic Explanations for Neural Ranking Models. In: Proceedings of ICTIR 2021. pp. 13–22. ACM.

[16] Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: Computational Argumentation Quality Assessment in Natural Language. In: Proceedings of EACL 2017. pp. 176–187, http://aclweb.org/anthology/E17-1017

[17] Yates, A., Arora, S., Zhang, X., Yang, W., Jose, K.M., Lin, J.: Capreolus: A Toolkit for End-to-End Neural Ad Hoc Retrieval. In: Proceedings of WSDM 2020. pp. 861–864. ACM. https://doi.org/10.1145/3336191.3371868