Webis at TREC 2018: Common Core Track

Alexander Bondarenko[1] Michael Völske[2] Alexander Panchenko[3]
Chris Biemann[3] Benno Stein[2] Matthias Hagen[1]

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

[2]Bauhaus-Universität Weimar
first.last@uni-weimar.de

[3]Universität Hamburg
last@informatik.uni-hamburg.de

ABSTRACT

This paper gives a brief overview of the Webis network's participation in the TREC 2018 Common Core track. The basic idea applied in our approach is to axiomatically re-rank the top-50 results of BM25F for those topics that seem to be argumentative. To this end, we use three axioms with the goal of covering some aspects of argumentativeness in text documents. If all three argumentative axioms favor a re-ordering of two documents, they "overrule" the initial ranking and the documents change their ranks.

1 INTRODUCTION

Our idea of the argumentative re-ranking for argumentative queries (i.e., queries for which results containing good argumentation might be the most promising ones) basically follows our last year's Common Core track approach [5]: capture preferences for more argumentative documents using ideas of axiomatic re-ranking [6]. Compared to last year, we did change the axioms and the weighting scheme a bit and only apply the re-ranking to the 25 topics / queries manually judged as argumentative .

2 WEBIS COMMON CORE TRACK RUNS

We briefly describe our procedure for identifying argumentative queries / topics and give a brief explanation of the axioms and their weights that we have selected to re-rank the BM25F top-50 results for argumentative queries.

2.1 Identifying Argumentative Queries

We manually went through the topic titles of this year's Common Core track and labeled the ones as argumentative where our impression was that some user submitting the title as a query might look for arguments in the resulting retrieved documents. Only for those topics that were labeled as argumentative, the axiomatic re-ranking is invoked to favor "argumentative" results. This is different to our last year's TREC Common Core track runs where we simply applied re-ranking with two argumentative axioms to each topic. The 25 topics from the TREC 2018 Common Core track that we labeled as potentially argumentative are given in Table 1.

2.2 Argumentative Axioms

One idea of axiomatic thinking for information retrieval [1, 4] is to identify axioms (i.e., constraints) that good retrieval models should fulfill. In a previous axiomatic re-ranking study, we have shown that weighted combinations of "all" known axioms can improve the initial ranking of several baseline retrieval systems [6]. Our last year's TREC approach then focused on two rather simplistic axioms that should capture argumentativeness during the re-ranking. However, re-ranking the results for all queries with argumentative axioms did not really show any improvements. We thus suggest the following three more fine-grained axioms in our this year's approach and only selectively apply the re-ranking to the top-50 BM25F results to potentially argumentative topics / queries.

Axiom ArgUC (Argumentative Units Count). The general idea of the ArgUC axiom is to favor documents that contain more argumentative units (i.e., a document is heuristically viewed as more "argumentative" when it contains more arguments).

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

Axiom QTArg (Query Term Occurrence in Argumentative Units). The general idea of the QTArg axiom is to favor documents where the query terms appear closer to argumentative units (i.e., a document is heuristically viewed as more argumentative on the query topic when the query terms appear in argumentative units instead of non-argumentative units).

Formalization. Let Q be an argumentative query consisting of the single term q, D_1 and D_2 be two retrieved documents, and Arg_D be the set of argumentative units of a document D. If $length(D_1) \approx_{10\%} length(D_2)$ and if $q \in A_{D_1}$ 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 [7, 10], 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 be an argumentative query consisting of the single term q, D_1 and D_2 be two retrieved documents, and $1^{st}position(q, Arg_D)$ be the first position in an argumentative unit of document D where the term q appears. If $length(D_1) \approx_{10\%} length(D_2)$ and if $1^{st}position(q, Arg_{D_1}) < 1^{st}position(q, Arg_{D_2})$ then $rank(D_1, Q) > rank(D_2, Q)$.

Similar to our general axiomatic re-ranking pipeline [6], in practice, we relax the axioms QTArg and QTOArg a bit more to also cover multi-term queries.

TREC 2018, November 14–16, 2018, Gaithersburg, Md.
2018. Webis network [webis.de].

2.3 Argumentative Unit Detection

To detect the argumentative units of a document, we use an own extension[1] of the system based on the BiLSTM-CNN-CRF neural network model developed by Reimers and Gurevych [8] that we retrained on the essay dataset created by Stab and Gurevych [9] employing fastText.cc vectors proposed by Bojanowski et al. [3] as representations.

2.4 Axiom Weights

Following our general axiomatic re-ranking pipeline [6], we employ an axiom ORIG in addition to the three argumentative axioms. The ORIG axiom simply corresponds to the preferences induced by the baseline retrieval system's ranking—BM25F in our case. The four different axioms' are being weighted for linearly combining the respective preference matrices [6]. As for our TREC 2018 Common Core track approach, we set the weight of ORIG to 0.43 and the argumentative axioms' weights to 0.19. The underlying idea is that the argumentative axioms have equal weights and can only "overrule" an ORIG preference iff they all agree—a later fine-tuning of the weights might be a promising direction.

2.5 The Webis Runs

The result lists of our three runs submitted to the TREC 2018 Common Core track do not fully reflect the above described argumentative idea, though. Our main goal with our three runs was to gather judgments for the whole top-50 of the BM25F baseline assuming a pooling depth of 20 would be used for the assessment. We wanted to later be able to experiment with other axiomatic re-ranking schemes for the top-50 results of BM25F without having to deal with a lot of documents without judgments.

We simply used the topic titles without any further processing or expansion as queries to an Elasticsearch BM25F index with the fields *title* (the document title; weight: 3) *summary1* (the first three sentences of the document body; weight: 2), *summary2* (TextRank-based summarization [2]; weight: 2), and *content* (document body; weight: 1), while the BM25 parameters were set to the "default" $k_1 = 1.2$ and $b = 0.75$. An optimization of these parameters with respect to the Washington Post corpus, application of some query expansion, etc. are natural possible optimization steps. Our run *webis-argument* has as its top-20 results the top-20 of BM25F for non-argumentative topics and for argumentative topics it has the top-20 results of the argumentative re-ranking of the BM25F's top-50 results. This way, *webis-argument* is the representative of our general idea. In the *webis-bm25f* and the *webis-baseline* runs, though, we used the remaining top-50 documents returned by BM25F not included in the *webis-argument* run + some random 10 documents from the BM25F's initial ranks 51 to 100 as the respective top-20 ranks.

2.6 Evaluation

We analyze the performance of the re-ranking on the argumentative topics in more detail. Table 1) reports the nDCG@10 of the BM25F baseline and the nDCG@10 of the argumentative axiom re-ranking for the 25 argumentative topics (manually labeled as such before submitting our runs to TREC, cf. Section 2.1). The argumentative re-ranking improves upon the baseline by an nDCG@10-change of at least 0.05 for four topics (bold in Table 1) but reduces it for six topics; the average nDCG@10 of 0.64 does not change.

Table 1: The 25 manually labeled argumentative topics from the TREC 2018 Common Core track ordered by the BM25F baseline's nDCG@10. The nDCG@10 of the argumentatively re-ranked top-50 results (and the difference to the baseline) are also given (the four topics with argumentative improvements ≥ 0.05 in bold).

Topic	Title / Query	BM25F	Axiom. Re-Ranking
801	africa polio vaccination	1.00	1.00
646	food stamps increase	1.00	1.00
824	bezos purchases washington post	0.99	0.99
814	china one-child impact	0.97	0.98 (+0.01)
445	women clergy	0.97	0.92 (-0.05)
803	**declining middle class in u.s.**	0.91	0.98 (+0.07)
375	hydrogen energy	0.91	0.88 (-0.03)
802	women driving in saudi arabia	0.83	0.81 (-0.02)
806	computers & paralyzed people	0.82	0.82
378	**euro opposition**	0.81	1.00 (+0.19)
805	eating invasive species	0.77	0.81 (+0.04)
809	protect earth from asteroids	0.76	0.67 (-0.09)
818	eggs in a healthy diet	0.66	0.70 (+0.04)
816	federal minimum wage increase	0.63	0.46 (-0.17)
690	college education advantage	0.63	0.50 (-0.13)
347	wildlife extinction	0.54	0.53 (-0.01)
321	women in parliaments	0.53	0.49 (-0.04)
341	**airport security**	0.52	0.72 (+0.20)
813	marijuana potency	0.50	0.43 (-0.07)
400	amazon rain forest	0.50	0.30 (-0.20)
426	**law enforcement, dogs**	0.43	0.63 (+0.20)
812	social media & teen suicide	0.31	0.30 (-0.01)
810	diabetes & toxic chemicals	0.00	0.00
393	mercy killing	0.00	0.00
350	health & computer terminals	0.00	0.00

3 CONCLUSION

We have used three basic axioms that aim to capture some rough ideas about documents' argumentativeness to re-rank results for queries that seem to be "argumentative" (i.e., that may benefit from argumentative results). Our first inspections of our runs' results show some promising effects of improved performance for some topics. We will try to fine-tune the detection of the argumentative units and axiom weights but we will also aim at developing axioms capturing further different angles of argumentativeness.

Natural other next steps are to improve the weighting scheme through larger-scale training, and to better detect argumentative topics (e.g., taking into account a topic's description or narrative for the labeling and / or looking at results judged as relevant more deeply for the Core tracks or past TREC tracks that also used newspaper collections).

Also interesting could be more fine-grained pre-conditions for the axioms besides the current similar-length constraint. An example could be to only try to axiomatically re-rank documents if their original retrieval score is rather similar.

[1] Available at http://ltdemos.informatik.uni-hamburg.de/argsearch/.

ACKNOWLEDGMENTS

This work has been partially supported by the Deutsche Forschungsgemeinschaft (DFG) within the project “Argumentation in Comparative Question Answering (ACQuA)” (HA 5851/2-1 and grant BI 1544/7-1) that is part of the Priority Program “Robust Argumentation Machines (RATIO)” (SPP-1999).

We also thank Oleksiy Oliinyk and Philipp Heidenreich from Universität Hamburg for providing the pre-trained models and a RESTful API that we used for argumentative unit detection (cf. Section 2.3).

REFERENCES

[1] E. Amigó, H. Fang, S. Mizzaro, and C. Zhai. Axiomatic Thinking for Information Retrieval: And Related Tasks. In *Proceedings of SIGIR 2017*, pages 1419–1420.

[2] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer. Variations of the Similarity Function of TextRank for Automated Summarization. *arXiv:*, 1602.03606.

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching Word Vectors with Subword Information. *arXiv*, 1607.04606.

[4] H. Fang, T. Tao, and C. Zhai. A Formal Study of Information Retrieval Heuristics. In *Proceedings of SIGIR 2004*, pages 49–56.

[5] M. Hagen, Y. Ajjour, J. Kiesel, P. Adineh, and B. Stein. Webis at TREC 2017: Open Search and Core Tracks. In *Proceedings of TREC 2017*.

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

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

[8] N. Reimers and I. Gurevych. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of EMNLP 2017*, pages 338–348.

[9] C. Stab and I. Gurevych. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659, 2017.

[10] A. D. Troy and G. Zhang. Enhancing Relevance Scoring with Chronological Term Rank. In *Proceedings of SIGIR 2007*, pages 599–606.