Quality-aware Argument Retrieval with Topical Clustering

Notebook for the Touché Lab on Argument Retrieval at CLEF 2021

Lukas Gienapp[1]

[1]Leipzig University, Leipzig, 04109, Germany

Abstract

We present a specialized approach to argument retrieval which combines both the general argumentative quality of texts, as well as the latent semantic (topic-)space of the document collection as boost factors to a general-purpose retrieval model to address the specific domain requirements of argument search. This setup aims to satisfy our three hypothesized aspects of an argumentative information need: quality-aware result ranking, near-complete topical coverage, and text proximity to the query.

Keywords

information retrieval, argument retrieval, argument quality, latent semantic clustering, CEUR-WS

1. Introduction

Searching the web, where information on virtually any topic can be accessed, has become a highly influential factor in everyday decision making. However, in many cases, an information need can be presumed that is addressed best not by single correct answer, or an unfiltered list of similar documents, but by a faceted view of different aspects of the search topic at hand. To this end, traditional approaches to web search only serve a diminished purpose, hence why specialized retrieval systems for this domain have to be developed, generating insights that support the user in forming well-justified opinions.

The first task of the Touché Shared Task Bondarenko et al. [1] supports such everyday decision making by incentivizing the development of specialized systems for argument retrieval for controversial questions. The aim of such systems is to retrieve argumentative texts relevant to controversial topics of general societal interest, which should be useful in conversations, debates, or forming an individuals' opinion on the topic at hand. In this paper, we contribute such a retrieval system, based on three hypothesized aspects of an argumentative information need: quality-aware result ranking, near-complete topical coverage, and text proximity to the query.

In contrast to established general-purpose retrieval models, our proposed method therefore does not only rank by term proximity to the query using the general-purpose Dirichlet language model for retrieval, but additionally takes into account the argumentative quality of text snippets,

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

lukas.gienapp@uni-leipzig.de (L. Gienapp)

0000-0001-5707-3751 (L. Gienapp)

as estimated using a support vector regression model, and the latent semantic space of the document collection, calculated by performing clustering on phrase embeddings. In Section 2, we review existing approaches to argument retrieval, and derive our three information need facets from related work. In Section 3, we introduce our method and provide detailed information on each of the components of our argument retrieval model. Section 4 provides first insight into the models' performance, while Section 5 gives concluding remarks.

2. Related Work

This section provides an overview of work that focuses specifically on the retrieval and ranking of argumentative documents. Throughout, we assume a static document collection, namely the args.me corpus [2], comprised of 387,740 arguments crawled from online debate portals. Based on this, Section 2.1 describes several existing retrieval approaches. Section 2.2 reviews different viewpoints on what an argument search system should achieve, influencing the design decision made throughout this paper.

2.1. Argument Retrieval Models

Bondarenko et al. [3] identify three central components of an argument retrieval system, based on a review of systems submitted to the Touché Shared Task: (1) an initial retrieval strategy; (2) an augmentation component, where results are extended by either expanding the query set, or directly based on documents features in the initially retrieved document set; (3) a (re)ranking component based on a primary document feature, influencing the final document scoring. We structure the literature review around each of these components, drawing inspiration for our own system at each step.

Initial Retrieval. As one of the first publicly available systems focusing on argument search, Wachsmuth et al. [4] present Args[1], implementing a fulltext search engine based on the args.me corpus utilizing the Okapi BM25 retrieval model [5]. In addition to BM25, Potthast et al. [6] evaluate three other general retrieval models for argument search, taking argument quality into account besides relevance, and find the DirichletLM model [7] performing best on average. This is corroborated by Bondarenko et al. [3], where the stock DirichletLM baseline system placed among the top systems evaluated. Dumani and Schenkel [8] use a parameter-free divergence-from-randomness model for initial retrieval in their pipeline, yet they do not provide an ablative evaluation characterizing the baseline performance of this step only. Beyond traditional retrieval models, employing large transformer-based language models for argument search has been successfully demonstrated by Akiki and Potthast [9], who use 512-dimensional phrase embeddings produced by the Universal Sentence Encoder (USE) [10] to calculate query proximity. Beyond argument search, the USE has been applied to general information retrieval [11] and numerous NLP tasks [10].

[1] www.args.me. Unless otherwise noted, all URLs in this paper have been last accessed on June 29, 2021 and were archived in the Wayback Machine

Result Augmentation. The result augmentation step aims at adding arguments to the result set that were not identified by the initial retrieval. One particular method of achieving such augmentation, first applied by Boltuzic and Snajder [12], is clustering—the general idea being to include all arguments that are members of the same (precomputed) clusters as documents already present in the initial results. Dumani and Schenkel [8] exploit the dual structure of the args.me corpus and group together arguments that share an identical claim. A notable shortcoming here is the strict identity of conclusions as clustering criterion, possibly leading to very small groupings. Dumani et al. [13] improve on this, utilizing phrase embeddings as calculated by models like Sentence-BERT [14], or InferSent [15] to project argument snippets into a clusterable vector space. Akiki and Potthast [9] follow a similar approach, using KMeans clustering on USE-embeddings to obtain semantic clusters of arguments in the args.me corpus.

Reranking. Since the result augmentation introduces a heap of previously not considered arguments into the result set, a reranking is warranted to expand the scoring beyond intial query similarity. In one of the top-scoring system at the first Touché Shared Task, Bundesmann et al. [16] propose argument quality as reranking feature, predicted using support vector regression. Other proposed reranking features include sentiment scoring [17], author credibility [3], and readability [3], but all to only limited success.

2.2. Considerations on the goals of Argument Search

To provide a motivation for the design choices made in Section 3, we consider different aspects of what a useful argument search system should provide. The underlying assumption here is that is system is used for conversational argument search: it is to provide assistance to users collecting argumentative evidence on various societal topics, to either provide debate assistance, or fulfill a personal informational need [3]. Besides this general goal, more specific requirements are placed on an argument search system that extend beyond general information retrieval. Following the propositions made by Wachsmuth et al. [18] and [4], Potthast et al. [6] argue that the evaluation of argument retrieval models should not only incorporate the classic evaluation criterion of relevance, but also include argument quality as additional evaluation feature, as differences between relevance-oriented effectiveness and quality-oriented effectiveness can be observed. This in turn means that argument search should maximize not only the relevance, but also the argumentative quality of its results.

Another issue which is partly raised by Boltuzic and Snajder [12] is the wording of argumentative text. They observe language variability, i.e. the same abstract argument can be expressed in nearly infinitely many ways, which may lead to shortcomings for the retrieval quality of term-based ranking models. The authors tackle this issue by applying semantic clustering. Bundesmann et al. [16] further comment on result diversity, and integrate a measure of heterogeneity to increase the diversity of viewpoints within their top-ranked results. This diversity can be related to a cluster-based retrieval as well, as one cluster may contain many different and diverse viewpoints for a particular topic. Therefore, a topic-aware ranking model might also yield improved results.

3. Methodological Approach

Our method for argument retrieval is composed of three components, integrating the notions of (1) *textual relevance* (Section 3.1), i.e., the relevance as indicated by a term-frequency based retrieval model; (2) *topical relevance* (Section 3.2), i.e., the relevance as indicated by a semantic space, independent of term occurrences; and (3) *argumentative relevance* (Section 3.3), i.e., the argumentative quality of the results. This is similar to the three steps described by Bondarenko et al. [3]: *textual relevance* is akin to the initial retrieval, *topical relevance* relates to the result augmentation step, and *argumentative relevance* can be seen as reranking feature. However, the critical difference here is that we do not model these components as successive steps in a retrieval pipeline, but rather as complementary parts of a final relevance score.

3.1. Textual Component

The textual component is modeled by a classic and domain-independent information retrieval model relying on term statistics of documents to infer the proximity, i.e., potential relevance, of each document to the text query. Given the popularity and very favorable performance of the DirichletLM retrieval model in the prior Touché Shared Task, we rely on it to calculate textual relevance scores. We use the Lucene implementation of the DirichletLM model[2], which closely follows the original paper [7].

3.2. Topical Component

We embed all argument conclusions in the document collection into a 512-dimensional vector space using the Universal Sentence Encoder [10]. We choose this embedder over other phrase embedding models due to its widespread application, high usability, favorable performance, and previous usage in the field of argument search [9]. While Akiki and Potthast [9] use USE-embeddings of the complete argument texts to perform exhaustive nearest-neighbor lookup for individual arguments at retrieval time, we instead utilize the embedding vectors to perform KMeans-Clustering on only the arguments' conclusions, to allow for coherent clusters of topically similar arguments. The feasibility of this approach has been demonstrated by Akiki and Potthast [9], who conduct a similar clustering approach to verify the accuracy of their embedding space and find that the clusters obtained are both coherent (syntactically and semantically) and meaningful (encompassing specific topics). After clustering the conclusion space with $k = 300$ (a parameter choice that yielded accurate results upon manual review) each argument is associated with its clusters' centroid. Each argument is scored by cosine similarity between query and its cluster centroid. The centroid is chosen over the individual proximity of arguments to equally boost the ranking score of all arguments in a cluster/topic, which enables to rank arguments within a topic by a secondary feature, such as quality.

3.3. Argumentative Component

We follow the considerations of Bundesmann et al. [16] who predict argument quality using a support vector regression (SVR) model. While they note that reliable quality prediction is difficulty to achieve, the overall retrieval effectiveness achievable by incorporating such

[2] https://lucene.apache.org/core/8_7_0/core/org/apache/lucene/search/similarities/LMDirichletSimilarity.html

predictions is still sufficiently high. To improve on their method, we introduce a classification step prior to the quality prediction, which decides whether a given text span is argumentative or not; non-arguments are then automatically assigned the minimal quality score, while arguments are passed on to the predictor to infer a rating for argumentative quality. We train both the argumentative classification and the quality prediction model using the *Webis ArgQuality 20* dataset [19]. It contains argumentative quality ratings and a binary classification whether a text is an argument or not for a subset of 1,271 arguments from the args.me corpus. We rescale quality scores to a range of $[0, 1]$, and convert arguments to lowercase, remove english stopwords and vectorize them as TF/IDF vectors.

First, a support vector machine (SVM) is used for a binary classification to determine if a sample is an argument or not. Then, valid arguments receive their quality rating as estimated by a SVR model. The classification is trained on the complete (binary label) data, while the SVR is trained only the argument subset. Both models are evaluated with 10-fold cross-validation. The SVM classificator is achieves F-1 score of 0.88. The SVR regression model achieves a mean square error (MSE) of 0.1949. Both models can thus be deemed reasonably accurate. The combined model is then applied to predict a quality score for each premise contained in the args.me corpus. Texts classified as non-arguments receive a score of 0, while others receive their quality prediction as given in the 0-1 range.

3.4. Final Scoring

Given the three components of our retrieval system described above, the final score $S_{(q,d)}$ of a document d for a query q is given by

$$S_{(q,d)} = R(q, d) \cdot (1 + \omega_C \cdot C(q, d)) \cdot (1 + \omega_Q \cdot Q(d)) \tag{1}$$

where $R(q, d)$ is the initial document relevance score as produced by the DirichletLM model, $C(q, d)$ being the cosine distance between the query embedding and the topical cluster centroid d is associated with, and $Q(d)$ is the predicted quality score for d (independent of the query). $\omega_{C,Q}$ are weighting factors to fine-tune the model. Both $C(q, d)$ and $Q(d)$ are in $[0, 1]$ and thus boost the initial score, but never decrease it.

4. Evaluation

We implement the described method as an Elasticsearch-based retrieval system. Quality ratings as well as cluster centroids are pre-computed for efficient retrieval. At retrieval time, all documents in the collection are scored by DirichletLM, and for all the cosine distance from centroid to USE-embedded query is calculated. Documents are then ranked by the scoring formula in Equation 1. We submit five runs for evaluation, differing in the applied weighting factors. All setups are summarized in Table 1. The first three weighting schemes are used to test if a higher, lower or equal influence of topical relevance and quality ratings are beneficial. With the last two setups, we fix the initial relevance score at 1 to investigate whether the topical relevance alone (since it is query-dependent, as opposed to quality), can provide meaningful and accurate results, without depending on a term-frequency based model at all. To generate and submit runs for Touché 2021, the Tira platform was used [20].

Table 1
Weighting schemes and resulting nDCG scores for both relevance- and quality-based evaluation. Maximum per column marked in bold.

ω_c	ω_q	nDCG@5 (Relevance)	nDCG@5 (Quality)	Remark
10.0	5.0	**0.645**	0.839	
10.0	10.0	0.639	**0.841**	
5.0	10.0	0.637	0.833	
Touché Dirichlet Baseline		0.626	0.796	
0.1	5.0	0.004	0.767	$R(Q, D) = 1$
0.01	5.0	0.000	0.749	$R(Q, D) = 1$

The first three runs, enabling the Dirichlet-based textual component, show strong overall performance. For relevance-based evaluation, the added topical component yields a net increase in ranking performance compared to the Dirichlet-only Touché baseline. The ranking performance also correlates with parameter choice for w_c, as higher value results in higher nDCG@5. Overall, for relevance, our best approach places 9th among teams. For quality-based evaluation, the same trend can be observed: the quality-based scoring factor has tremendous impact on improving the argumentative quality of the results. Once again, the higher choice of w_q results in the higher ranking performance, however, only in conjunction with a high value of w_c as well. In terms of quality evaluation, the three approaches place first among all runs submitted to Touché. The two-stage prediction model can thus be deemed highly effective.

The latter two approaches, where the Dirichlet-based textual component has been turned off turn out to be unusable in practice. With an nDCG score of zero (for relevance), they provide effectively no use to a user. One possible reason for this is that the embedding space was constructed on arguments' conclusions only, which is not sufficient to ensure relevant search results. However, regarding argumentative quality, the system yields acceptable results, too.

5. Conclusion

We proposed a new approach to argument retrieval, combining several parts of existing systems that have shown favorable performance prior. The retrieval model is centered around three components: a classic term-frequency-based retrieval model (DirichletLM) and two boosting factors, incorporating topical relevance as indicated by a semantic clustering of the underlying data, and a quality prediction model. The approach can be deemed successful. For both relevance and quality as evaluation dimensions, the system yields useful results. For quality, it places highest among the participants of this years' Touché lab. The evaluation has also shown room for future improvements: specifically the topical component performs sub-par, and needs to be revisited. Extending the embeddings to not only include conclusions, but also premises, maybe even in terms of a dual embedding space promises better results. Parameter fine-tuning for the Dirichlet model also promises an increase in ranking performance and will be made possible by the increased availability of relevance judgements from this years' iteration of Touché.

References

[1] A. Bondarenko, L. Gienapp, M. Fröbe, M. Beloucif, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2021: Argument Retrieval, in: D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval. 43rd European Conference on IR Research (ECIR 2021), volume 12036 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2021, pp. 574–582. doi:10.1007/978-3-030-72240-1_67.

[2] Y. Ajjour, H. Wachsmuth, J. Kiesel, M. Potthast, M. Hagen, B. Stein, Data Acquisition for Argument Search: The args.me corpus, in: C. Benzmüller, H. Stuckenschmidt (Eds.), 42nd German Conference on Artificial Intelligence (KI 2019), Springer, Berlin Heidelberg New York, 2019, pp. 48–59. doi:10.1007/978-3-030-30179-8_4.

[3] A. Bondarenko, M. Fröbe, M. Beloucif, L. Gienapp, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2020: Argument Retrieval, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes Papers of the CLEF 2020 Evaluation Labs, volume 2696 of *CEUR Workshop Proceedings*, 2020.

[4] H. Wachsmuth, M. Potthast, K. Al-Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, B. Stein, Building an Argument Search Engine for the Web, in: K. Ashley, C. Cardie, N. Green, I. Gurevych, I. Habernal, D. Litman, G. Petasis, C. Reed, N. Slonim, V. Walker (Eds.), 4th Workshop on Argument Mining (ArgMining 2017) at EMNLP, Association for Computational Linguistics, 2017, pp. 49–59.

[5] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: D. K. Harman (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, volume 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1994, pp. 109–126.

[6] M. Potthast, L. Gienapp, F. Euchner, N. Heilenkötter, N. Weidmann, H. Wachsmuth, B. Stein, M. Hagen, Argument Search: Assessing Argument Relevance, in: 42nd International ACM Conference on Research and Development in Information Retrieval (SIGIR 2019), ACM, 2019. doi:10.1145/3331184.3331327.

[7] C. Zhai, J. D. Lafferty, A study of smoothing methods for language models applied to information retrieval, ACM Trans. Inf. Syst. 22 (2004) 179–214. doi:10.1145/984321.984322.

[8] L. Dumani, R. Schenkel, Quality-aware ranking of arguments, in: M. d'Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, ACM, 2020, pp. 335–344. doi:10.1145/3340531.3411960.

[9] C. Akiki, M. Potthast, Exploring Argument Retrieval with Transformers, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes Papers of the CLEF 2020 Evaluation Labs, volume 2696, 2020.

[10] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, R. Kurzweil, Universal sentence encoder, CoRR abs/1803.11175 (2018). URL: http://arxiv.org/abs/1803.11175. arXiv:1803.11175.

[11] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Ábrego, S. Yuan, C. Tar, Y. Sung, B. Strope, R. Kurzweil, Multilingual universal sentence encoder for semantic

retrieval, in: A. Çelikyilmaz, T. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 87–94.

[12] F. Boltuzic, J. Snajder, Identifying prominent arguments in online debates using semantic textual similarity, in: Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA, The Association for Computational Linguistics, 2015, pp. 110–115. doi:10.3115/v1/w15-0514.

[13] L. Dumani, C. K. Kreutz, M. Biertz, A. Witry, R. Schenkel, Segmenting and clustering noisy arguments, in: D. Trabold, P. Welke, N. Piatkowski (Eds.), Proceedings of the Conference "Lernen, Wissen, Daten, Analysen", Online, September 9-11, 2020, volume 2738 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 23–34.

[14] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. doi:10.18653/v1/D19-1410.

[15] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised learning of universal sentence representations from natural language inference data, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, Association for Computational Linguistics, 2017, pp. 670–680. doi:10.18653/v1/d17-1070.

[16] M. Bundesmann, L. Christ, M. Richter, Creating an argument search engine for online debates, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[17] C. Staudte, L. Lange, Sentarg: A hybrid doc2vec/dph model with sentiment analysis refinement, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[18] H. Wachsmuth, N. Naderi, I. Habernal, Y. Hou, G. Hirst, I. Gurevych, B. Stein, Argumentation quality assessment: Theory vs. practice, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers, Association for Computational Linguistics, 2017, pp. 250–255. doi:10.18653/v1/P17-2039.

[19] L. Gienapp, B. Stein, M. Hagen, M. Potthast, Efficient pairwise annotation of argument quality, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 5772–5781.

[20] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.