

Exploring Document Expansion for Argument Retrieval

Notebook for the Touché Lab on Argument Retrieval at CLEF 2021

Alina Mailach, Denise Arnold, Stefan Eysoldt and Simon Kleine

Leipzig University, Augustusplatz 10, 04109 Leipzig, Germany

Abstract

Processing of opinion based information is an increasingly relevant task in modern times. Especially in regards to complex and morally ambiguous topics, users search for high-quality information which leads to the necessity of automatically processing information of argumentative nature. This notebook documents our attempt to improve argument retrieval using expansion methods for documents as a contribution to Touché@CLEF 2021 as team Hua Mulan. Before runtime we expand arguments by predicting queries and hallucinating arguments using Transformer architectures and a more computational efficient approach based on TF-IDF. Compared to ad-hoc retrieval of the original args.me corpus with Dirichlet Language Model argument hallucination improved the baseline when evaluated on argument quality, while no improvements were obtained when argument relevance was evaluated.

Keywords

Argument retrieval, document expansion, query prediction

1. Introduction

The rapid digitalization and development of novel technologies has led to an unprecedented amount of information, that has to be processed by individuals and transforms our society accordingly. Subsequently, this affects especially the ways in which we debate and form opinions – this holds true for simple decision-making as well as for morally ambiguous topics and politics. A great challenge in this respect is the accurate and automatic identification, validation and retrieval of argumentative patterns in order to help users deal with the tremendous amount of information and ease opinion formation processes.

The shared Task Touché@CLEF 2021 [1, 2] is the first shared Task focusing on argument retrieval. Task 1 is dedicated to developing methods to identify and score conversational arguments in a search scenario, in which the user tries to find good arguments regarding a relevant, ambiguous topic. In this notebook we describe our findings as Team Hua Mulan after evaluating document expansion methods inspired by approaches in regular information retrieval on the task of retrieving arguments from the args.me corpus [3].

Our work builds upon several contributions to the Touché@CLEF shared Task 1 in 2020 [4]. Closely related but distinct contributions are the query expansion methods using transformers


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

✉ mailach@informatik.uni-leipzig.de (A. Mailach); arnold@studserv.uni-leipzig.de (D. Arnold);

se57nafy@studserv.uni-leipzig.de (S. Eysoldt); pge12kaa@studserv.uni-leipzig.de (S. Kleine)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

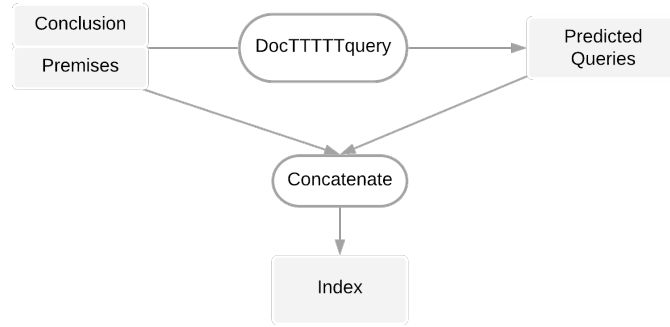


Figure 1: Procedure of document expansion using docTTTTTquery

by Akiki & Potthast [5] as well as query expansion using WordNet synonyms [6]. Our approach diverges on the moment of execution as well as on the subject to expansion. While both former contributions focus on expanding queries at runtime, we expand documents prior to indexing. While the first two expansion methods are based on the Transformer architecture, the third one is a more intuitive approach based on finding synonyms. In the following sections we further elaborate on our approaches and findings.

2. Document Expansion

A gap that many retrieval approaches try to bridge is the issue of mismatch between terms in a query and terms in documents relevant to this query. This mismatch is caused by different words describing the same content. A possibility to raise the probability of the retrieval of a document which in it's original form does not contain the keywords in the query, is the enrichment of documents with terms that are not yet contained, but are very likely to be contained by a query that is used to search for this documents. The following part is dedicated to describe the three approaches. This section is describing the implementation and evaluation of expanding documents by predicting relevant queries (2.1), hallucination of arguments (2.2) and extracting synonyms (2.3).

2.1. Query Prediction

Predicting queries and augmenting the documents with these predicted queries was introduced in 2019 by Nogueira et al. [7] and called *doc2query*. The approach is grounded on the idea of conceptualizing the retrieval process as a question-answering system, in which the query represents a question for which the user searches the right answer in order to satisfy her information need. The authors build their query-prediction system by training a vanilla sequence-to-sequence model on the MS Marco dataset [8]. In 2020 Nogueira et al. published an improved version, called *docTTTTTquery* [9], which's main difference to *doc2query* is the basement on a T5 (Text-to-Text Transfer Transformer) [10] encoder-decoder architecture which was also trained on approximately 500.000 passage-query pairs and is publicly available in the

Table 1
Queries predicted by docTTTTQuery

Premises	Conclusion	Predicted Queries
Teachers who perform below benchmarks such as retention, attendance, academic performance results, assessing required learning outcomes and student feedback, should not be allowed tenure because students suffer to be successful and colleges suffer in graduation rates.	Colleges should abolish the ability for teachers to be tenured.	<ul style="list-style-type: none"> • Why should teachers not be tenured? • Why should tenured teachers be banned? • Why should tenured teachers not be allowed to work at a college? • Why should tenure be abolished?

authors github repository¹.

For the expansion of the arguments we used the original pretrained model and predicted ten queries per argument. Figure 1 shows the general procedure of the prediction. Premises and conclusion were concatenated, serving as the input and tokenized with Huggingface’s tokenizer². The input tensors are truncated to 512 tokens and are subsequently fed to the model. After detokenization, the predicted queries were appended to the original premises. Just as in the original work[9], we did not indicate the expansion with any special characters. In Table 1 an example argument with the predicted queries is given. We expected the additional information added by the predicted queries to reduce the issue of term mismatch and therefore improve retrieval performance.

2.2. Argument Hallucination

Akiki & Potthast [5] explored query expansion scenarios using different Transformer methods. Running multiple text sequences generated by *Generative Pretrained Transformer 2* (GPT-2)³ against the index improved retrieval quality compared to a simple baseline. We adapted the idea of generating arguments in a way that fits document expansion. While Akiki & Potthast have generated 24 different sequences for each query, we generated two sequences for each of the 72173 unique conclusions in the args.me corpus. This should give premises belonging to relevant conclusions a boost in the retrieval process by adding words and thus diversifying the language model for those arguments.

Just as Akiki & Potthast we wrapped each conclusion in an interview-like scenario using hyphens to indicate the conversational nature of the text. We further augmented the conclusion with a positive or negative prompt, leading GPT-2 in either of those directions. Since the retrieval process and the metric is agnostic towards the stance of an argument, we expanded each

¹<https://github.com/castorini/docTTTTQuery>

²https://huggingface.co/transformers/model_doc/t5.html

³<https://huggingface.co/gpt2>

Table 2

Example arguments hallucinated by GPT-2

Negative prompt	Positive prompt
-What do you think of: <i>Colleges should abolish the ability for teachers to be tenured?</i> -	
The answer is no, because the current system of teaching in America has been a disaster since it began and continues today (see my post on this topic). The only way we can get rid from that situation would be by abolishing teacher tenure at all levels and I'm not talking just one level here; there are many more who have already done so as well! But if they don't want their children taught like other kids then why bother with them when your child will learn something new every day?! [...]	The answer is yes, because the College of Education has been abolished. The only way that it can continue as a college in this country and still exist today would be if we had an independent school system where all students were treated equally regardless what their academic abilities are or how they perform at work (and not just on campus). This means there wouldn't even have any problem with having one teacher who was able/unable get tenure from his job [...]

argument with the positively and negatively generated sequences. An exemplary generation can be observed in table 2. Note that we decided not to generate sequences with neutral prompts, since our experiments revealed no relevant information gain compared to the positive and negative versions while increasing runtime by a third.

2.3. Synonym Extraction

The former mentioned methods are using complex neural networks and are therefore heavily relying on computational capacities and hardware acceleration. When thinking about real-world retrieval scenarios, expanding all arguments prior to or at indexing time, would lead to computational and runtime issues. Finding a more basic strategy to argument expansion is therefore necessary and interesting. Thus we implemented an approach similar to Bundesmann et al. [6].

For each argument we extracted the main keywords identified using *term frequency-inverse document frequency* (TF-IDF), indicating that a term is occurring relatively often in a document, compared to the occurrence in the rest of the args.me corpus and therefore is more relevant than other words. We used the scikit-learn library⁴ for computing TF-IDF and then augmented the argument with synonyms. For each argument the top 10 keywords that appeared in a maximum of 20% of the documents in the corpus were extracted. Subsequently, we searched for synonyms in the WordNet database [11] and appended them to the original premises. On average we extracted 8.80 keywords per argument (with a standard deviation of 2.55) and added on average 25.30 synonyms (with a standard deviation of 11.70).

⁴www.scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

Table 3
Results

	Relevance		Quality	
	nDCG@5 _{mean}	CI _{95%}	nDCG@5 _{mean}	CI _{95%}
Baseline (Dirichlet)	0.626	[0.550, 0.691]	0.796	[0.755, 0.838]
Query Prediction	0.518	[0.446, 0.588]	0.654	[0.584, 0.724]
Argument Hallucination	0.620	[0.545, 0.685]	0.811	[0.770, 0.849]
Synonym Extraction	0.620	[0.549, 0.685]	0.789	[0.750, 0.830]

3. Evaluation

For the evaluation, each of the augmented corpora was indexed using Elasticsearchs built-in similarity based on *Dirichlet Language Model* (DirichletLM) to obtain the thousand most fitting arguments. DirichletLM was mainly chosen for performance reasons, as it proved to be most adequate for ad-hoc argument retrieval [12] and methods based on DirichletLM outperformed approaches based on other retrieval methods [4]. We ran non-systematic pre-tests on the corpus with no augmentation and found $m = 2148$ to retrieve good results. The approaches were evaluated on TIRA platform [13] for comparability and reproducibility. Table 3 shows mean results for relevance and quality of the retrieved arguments. In terms of retrieval, none of our approaches was able to improve baseline ad-hoc retrieval, while argument hallucination using GPT-2 achieved slightly higher results in argument quality. One reason for these results could be the expansion of all documents which leads to boosting less relevant documents. Further research could explore expansion of only high quality arguments to selectively improve retrieval of these documents.

4. Conclusion

We investigated the effect of different document expansion methods on argument retrieval. The examined methods tackled the issue of term mismatch using three different generative approaches. We used *docTTTTTquery* to predict relevant queries and hallucinating arguments using GPT-2. To test another approach for solving the information mismatch we extracted keywords and searched for synonyms in the WordNet corpus. The augmented corpora were indexed and retrieved using DirichletLM. Finally, none of the introduced approaches was able to beat the simple Baseline of ad-hoc retrieval using DirichletLM in terms of argument relevance. When evaluating argument quality, expanding documents using hallucinated arguments slightly improved retrieval.

References

- [1] 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. Névelol (Eds.), Working Notes Papers

of the CLEF 2020 Evaluation Labs, volume 2696 of *CEUR Workshop Proceedings*, 2020. URL: <http://ceur-ws.org/Vol-2696/>.

- [2] 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: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Working Notes Papers of the CLEF 2021 Evaluation Labs, CEUR Workshop Proceedings, 2021.
- [3] Y. Ajjour, H. Wachsmuth, J. Kiesel, M. Potthast, M. Hagen, B. Stein, Data Acquisition for Argument Search: The args.me corpus, in: C. Benz Müller, 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.
- [4] 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. Névél (Eds.), Working Notes Papers of the CLEF 2020 Evaluation Labs, volume 2696 of *CEUR Workshop Proceedings*, 2020. URL: <http://ceur-ws.org/Vol-2696/>.
- [5] C. Akiki, M. Potthast, Exploring Argument Retrieval with Transformers, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névél (Eds.), Working Notes Papers of the CLEF 2020 Evaluation Labs, volume 2696, 2020. URL: <http://ceur-ws.org/Vol-2696/>.
- [6] M. Bundesmann, L. Christ, M. Richter, Creating an argument search engine for online debates, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névél (Eds.), Working Notes Papers of the CLEF 2020 Evaluation Labs, volume 2696, 2020. URL: <http://ceur-ws.org/Vol-2696/>.
- [7] R. Nogueira, W. Yang, J. Lin, K. Cho, Document expansion by query prediction, CoRR abs/1904.08375 (2019). URL: <http://arxiv.org/abs/1904.08375>. arXiv:1904.08375.
- [8] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated machine reading comprehension dataset, in: T. R. Besold, A. Bor-des, A. S. d’Avila Garcez, G. Wayne (Eds.), Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016, volume 1773 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016. URL: http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.
- [9] R. Nogueira, J. Lin, A. Epistemic, From doc2query to docttttquery, Online preprint (2019).
- [10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: <http://jmlr.org/papers/v21/20-074.html>.
- [11] G. A. Miller, Wordnet: A lexical database for english, Commun. ACM 38 (1995) 39–41. URL: <https://doi.org/10.1145/219717.219748>. doi:10.1145/219717.219748.
- [12] 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. URL: <http://doi.acm.org/10.1145/3331184.3331327>.
- [13] 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.