Argument Search: Assessing Argument Relevance

Martin Potthast[1] Lukas Gienapp[1] Florian Euchner[2] Nick Heilenkötter[3]
Nico Weidmann[4] Henning Wachsmuth[5] Benno Stein[6] Matthias Hagen[7]

[1]Leipzig University [2]University of Stuttgart [3]University of Bremen [4]Karslruhe Institute of Technology
[5]Paderborn University [6]Bauhaus-Universität Weimar [7]Martin-Luther-Universität Halle-Wittenberg

ABSTRACT

We report on the first user study on assessing argument relevance. Based on a search among more than 300,000 arguments, four standard retrieval models are compared on 40 topics for 20 controversial issues: every issue has one topic with a biased stance and another neutral one. Following TREC, the top results of the different models on a topic were pooled and relevance-judged by one assessor per topic. The assessors also judged the arguments' rhetorical, logical, and dialectical quality, the results of which were cross-referenced with the relevance judgments. Furthermore, the assessors were asked for their personal opinion, and whether it matched the predefined stance of a topic. Among other results, we find that Terrier's implementations of DirichletLM and DPH are on par, significantly outperforming TFIDF and BM25. The judgments of relevance and quality hardly correlate, giving rise to a more diverse set of ranking criteria than relevance alone. We did not measure a significant bias of assessors when their stance is at odds with a topic's stance.

ACM Reference Format:
Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument Search: Assessing Argument Relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331327

1 INTRODUCTION

Decision making processes often come to a point where one is challenged by a *why*-question: a prompt to justify one's stance. An answer to a why-question may be a simple fact. More commonly, though, it requires the formulation of a justified claim: an argument.

The web is full of documents containing arguments such as news, blogs, discussions, or reviews. Still, leading search engines do not support the retrieval of arguments well. Search results on controversial topics are often riddled with populism, conspiracy theories, and one-sidedness. Though these may be popular argumentative techniques, they may not be exactly what one expects to be ranked high. Also, when users search for assistance on small life decisions in the form of arguments for or against a given option, the top search results do not reflect the argumentative landscape very well.

In both scenarios, stakeholders with vested interests in how a user decides compete for the top ranks through search engine optimization. Claiming the neutrality of their ranking algorithm and its robustness against exploitation, search engine operators understandably do not wish to take sides. Yet, simply letting the stakeholders compete for what opinions rank higher is not in the interests of a search engine's user base. Rather, search engines should identify all stakeholders and summarize their stances on a given argumentative topic to inform their users accordingly.

Argumentation analysis can be seen as an enabling technology in this respect. In the wake of recent advances in this field, the search for arguments (instead of the documents containing them) has been suggested. First prototypes of argument search engines are being developed and deployed both in industry (IBM's Project Debater [16]) and academia (args.me [22] and ArgumenText [19]). To allow for their rigorous evaluation, we contribute the first systematic user study on judging argument relevance alongside the first evaluation of four standard retrieval models (DirichletLM, DPH, BM25, TFIDF). In particular, we revisit the idea of relevance in the context of argument quality and shed light on the potential of assessor bias when judging controversial topics.[1]

2 RELATED WORK

Argument search is a new direction. The first system to retrieve arguments for a specified topic is IBM's Project Debater [16]—in order to compete with a human in a classical debate.[2] In 2017, Wachsmuth et al. [22] launched args.me, the first search engine for arguments on the web using a BM25 index of about 300,000 arguments crawled from online debating portals. Thereafter, ArgumenText [19], which retrieves argumentative sentences from the Common Crawl, and "multi-perspective answers" in the US version of Bing[3] have been published. Another loosely related case where argument search applies is the use of Wikipedia to debunk conspiracy theories on YouTube.[4] However, apart from some promising results of argument-based re-ranking in the TREC Common Core track [5], neither a comparison of suitable retrieval models for argument search nor a rigorous assessment of argument relevance has been carried out to date. We contribute the first TREC-style evaluation of four retrieval models on the 300,000 arguments indexed by args.me. In addition to the well-known information retrieval notion of relevance, we also assess other argument quality dimensions found in the argumentation literature (discussed in Section 3).

[1]All code and data is publicly available: https://github.com/webis-de/SIGIR-19
[2]https://www.research.ibm.com/artificial-intelligence/project-debater/
(All URLs have been archived on May 22, 2019, at the Internet Archive.)
[3]https://blogs.bing.com/search-quality-insights/february-2018/Toward-a-More-Intelligent-Search-Bing-Multi-Perspective-Answers
[4]https://youtube.googleblog.com/2018/07/building-better-news-experience-on.html

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '19, July 21–25, 2019, Paris, France
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331327

3 ARGUMENT SEARCH

Examining argument search more closely, we outline the commonalities and differences of existing technologies with respect to the information retrieval notions of retrieval task, retrieval model, relevance, and result presentation. The fundamental unit of retrieval in argument search are arguments. An argument is a claim accompanied by an arbitrary number of premises that justify the claim. A document may contain an arbitrary number of arguments.

Retrieval Tasks. In order to characterize the retrieval task, the central question is: Which information needs can an argument search engine cater to? These information needs, or use cases, can be derived from argumentation theory. There, argumentation is defined as a communicative process between parties, where either one party tries to win over the other (argument as *controversy* [9]), or where all parties collaborate to gain insights into an issue (argument as *debate* [17]). Mohammed's alternative formulation of identifying *intrinsic* and *extrinsic* argumentation goals [13] also supports this distinction: the former is about convincing an opponent of the acceptability of an opinion [20], and the latter is an inquiry to base decision making on [12]. In both cases, an argument search engine can contribute meaningful debate support.

In the context of the given argumentative use cases, the generic argumentation theory can be reinterpreted in terms of the classic dichotomy of ad-hoc retrieval versus task-based retrieval. At first glance, argument search appears rather to be of a task-based, exploratory nature, where a user needs to get to the bottom of a non-trivial issue by reviewing the argumentative discourse surrounding it. Such information needs can be called *deliberative*; they frequently arise during writing (an essay, blog post, thesis, speech etc.), when preparing for a discussion in an upcoming meeting, but also before purchase decisions. On closer inspection, ad-hoc retrieval plays a significant role in argument search, too: In text-based communication in both closed groups (e.g., WhatsApp) and open groups (Twitter), argument information is needed when being confronted with a claim. Attentive users want to double-check the claim (seeking *confirmation*), find a good reply or counterargument (*refutation*), and stand by their peers (*support*). The explicit use of a search engine may also be imagined as a system passively monitoring the discussion to prompt recommendations.

A daring next step then is to allow the argument retrieval system to actively enter into a conversation as an artificial social entity, for instance, enacting the role of the devil's advocate. If not for IBM's Project Debater and Google Duplex,[5] which recently demonstrated advanced conversational capabilities, this might have seemed more like a long-term goal for research and development. As it stands, however, these technologies pave the way for argument search to coalesce with conversational search, enabling informed debate at every dinner table through voice assistants.

Retrieval Models and Strategies. No retrieval models tailored to argument search have surfaced to date. The aforementioned search engines args.me and ArgumenText employ Lucene's BM25 model, whereas little is known about the models of Project Debater and Bing. Args.me and ArgumenText implement two diametrically-opposed retrieval pipelines: "mining-before-retrieval" (args.me) and "retrieval-before-mining" (ArgumenText). The mining-before-retrieval strategy presumes that argument mining is applied offline and that the extracted arguments are then indexed for later online retrieval. The retrieval-before-mining strategy uses a standard search engine to retrieve documents related to a given query. Argument mining is then applied on-the-fly on the top-ranked documents to extract and possibly re-rank arguments for the actual result page.

Mining-before-retrieval allows for more expensive offline indexing operations (e.g., computing argument PageRanks [23]), while this would slow down a retrieval-before-mining approach. The retrieval-before-mining strategy also hinges on the availability of suitable argument mining technology, whereas mining-before-retrieval can resort to distant supervision (as was done in collecting the args.me collection) to avoid flawed automatic argument mining results (sometimes observable in ArgumenText). Moreover, the retrieval-before-mining strategy relies on the retrieval model to retrieve argumentative documents—a different intermediate retrieval task, which either requires an additional focused retrieval approach, or scoring the documents' argumentativeness at indexing time. An advantage of retrieval-before-mining, however, is that it allows for deciding specifically for the query at hand whether a piece of text is argumentative with respect to that query or not.

Relevance and Quality. Wachsmuth et al. [21] recently surveyed the many argument quality dimensions distinguished in argumentation theory, organizing them within the widely accepted trichotomy of rhetorical, logical, and dialectical quality [4, 7, 25]. Rhetorical quality includes notions of persuasive effectiveness, correct language, vagueness, and style. An argument of high rhetorical quality is well-written and appealing to the audience. Logical quality refers to an argument's structure and composition. An argument of high logical quality is based on acceptable premises and combines them in a cogent way to support the argument's claim. Dialectical quality captures an arguments' contribution to the discourse. An argument of high dialectical quality is useful to support cooperative decision making or to resolve a conflict.

Argumentation theory also includes quality dimensions of so-called local and global relevance: Local (or probative) relevance refers to the logical composition of an argument, reflecting that the argument's premises provide support to its conclusion [3]; global (or dialectical) relevance captures that an argument contributes to a discussion [24]. The latter comes close to the notion of relevance prevailing in IR: it is a mixture of topic relevance and user relevance. A reconciliation of these different notions of relevance is beyond the scope of our paper, which is why we prefer to keep them separate for now, considering IR relevance alongside the more abstract concepts of rhetorical, logical, and dialectical quality.

Result Presentation. Retrieved arguments need to be presented in a suitable interface. Args.me and ArgumenText resemble the interfaces of traditional web search engines. Each argument is presented as a linked title and a text snippet. Additionally, args.me and ArgumenText also show pro/con labels to summarize an argument's gist. These labels, however, seem to be static and not to depend on the actual query. Understanding a user's stance from a query as being pro, con, or neutral and adapting a retrieved argument's label to the user's perspective is an interesting but difficult task for future improvement of argument search.

[5] https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

4 USER STUDY AND EVALUATION

In the absence of any evidence to the contrary, we work under the hypothesis that rhetorical, logical, and dialectical quality all play an important role in determining a ranking of relevant arguments. In this respect, we aim to assessing IR relevance and the three quality dimensions individually, which, in general, introduces new difficulties as per their subjective interpretation [11]. Though Habernal and Gurevych [8] argue that manual argument quality assessment can only be done relatively by comparing arguments, other works report on absolute quality assessments with reasonable agreement [15, 21]; we pursue the latter setup in our study.

Experimental Setup. We compare four standard IR models in a TREC-style setup on controversial topics, namely, the args.me system [22] (basically Lucene's BM25) and Terrier's [14] implementations of DPH [2], DirichletLM [26], and TFIDF [18]; all on the args.me collection. A G*Power analysis [6] yielded a least sample size of 16 topics to measure a presumed statistically significant performance difference between DPH (often performing well at TREC) and TFIDF. We hence selected 20 controversial issues, formulating a neutral and a biased topic per issue (40 topics in total); this allows the analysis of potential assessor bias. Table 1a shows an example.

At a pooling depth of $k = 5$, we recruited a different assessor for each topic to judge the pooled arguments in random order with respect to their topic relevance, plus their rhetorical, logical, and dialectical quality. Following other argument quality studies [21], a 4-point Likert scale (1 meaning low to 4 meaning high) was employed per assessment dimension; identified non-arguments received a score of -2 in all categories. The 40 assessors (31 male, 9 female; mean age 26, youngest 18, oldest 53), are volunteers from of a group of 170 students (plus 20 instructors), recruited at a summer academy of the German Academic Scholarship Foundation. A high personal integrity as well as strong interest in societal issues can be presumed. After the actual assessment, the assessors were asked about their personal stance on the topic assigned to them. In this respect, an independent survey of political orientation revealed that 80% vote for left-wing, green parties: an a-priori assessor bias.

Score distribution. Altogether, 437 arguments were judged across all topics (208 labeled as pro in the args.me corpus, 195 as con, 34 labeled as non-arguments by our assessors) resulting in the score distributions per quality dimension shown in Table 1b. The relevance scores are skewed towards the upper end, indicating that many highly relevant arguments were retrieved. Rhetoric and logic scores have a spike at 3, which might be explainable by the known reluctance of subjects to select extremes on Likert-based questions [1]. The uniform scores for dialectical quality may hint at an unclear explanation of that dimension in the survey; an additional "don't know" option might have helped.

Expert agreement. Wachsmuth et al. [21] ascertain that expert assessors are able to distinguish rhetorical, logical, and dialectical argument quality by showing that the assessments do not perfectly correlate on their set of arguments. Repeating the same analysis on the assessments of our set of arguments, we reproduce their results, showing that also lay assessors distinguish the three dimensions, and that a similar distribution of correlations is measured, thus mutually corroborating both results. Table 1d compares Pearson's ρ as measured for our assessments with those of the experts taken from [21, Table 3]. While our absolute scores are lower by 0.09 to 0.16, the relative relations are mostly the same. For instance, rhetorical quality correlates more with logical quality than with dialectical quality. In addition, by measuring the correlation of IR relevance against the other qualities, it seems that relevance has the most in common with dialectical quality.

Assessor bias. Since assessors were asked for their personal stance on biased topic versions, we can analyze the potential systematic bias of higher (or lower) average scores, dependent on whether the assessor (dis)agrees with the stance of an argument. The upper half of Table 1e shows the mean scores for every cross-category of assessor and argument stance. The hypothesis of a divergence between assessments on pro and con arguments is tested using the Mann-Whitney test; its p-values are shown in the lower half of the table. Category sizes are shown on the upper right of the table.

In short, no significant divergence pattern in the average scores by stances is apparent. However, we cannot determine whether a potential systematic bias is just masked by a data-inherent bias (e.g., pro arguments scoring higher on average in the logic dimension since pro arguments in the underlying args.me dataset are objectively "more logical"). At any rate, even if there was an (undetected) systematic bias, its effect is low—despite the left-wing orientation of the assessor cohort: when controlling for a topic's political orientation, assessor bias still remains insignificant.

Ranking performance. We measure the performance of the four retrieval models on the four dimensions as nDCG@5 [10]. The ideal ranking for a given topic is obtained by ordering its argument pool by descending relevance or quality score. Table 1f shows the obtained results with bootstrapped confidence intervals ($n = 10{,}000$, $\alpha = 0.95$). The observable differences of the confidence intervals (DirichletLM and DPH better than TFIDF which is relevance-wise better than args.me's BM25) were tested using a 1-way Anova test ($\alpha = 0.05$) that confirmed the visual impression: DirichletLM and DPH are rather indistinguishable but outperform the other two models on every category while the relevance difference of TFIDF and BM25 is the only other significant difference.

Regarding relevance, DirichletLM and DPH perform similar, and both outperform TFIDF and BM25, the latter of which performs worst. Also for the other three quality dimensions, DirichletLM and DPH perform similar, and both better than BM25 and TFIDF. Since DirichletLM and DPH achieve rather similar scores for all dimensions, one might think that they often retrieve the same arguments. However, the Jaccard index of the arguments they retrieve for all topics is just 38% (only for two topics, they return exactly the same arguments but rank them differently).

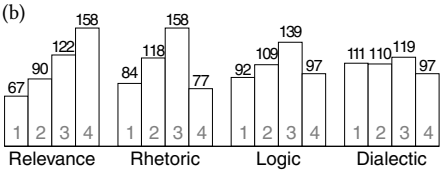
While DirichletLM and DPH are not really distinguishable via their mean nDCG scores on the four dimensions, DirichletLM has a lower variance. When looking at the absolute number of topics a model wins (Table 1c), DPH clearly is ahead in the relevance dimension while DirichletLM outperforms the others in 2 categories, being tied with DPH on rhetorical quality.

In the argument retrieval scenario that we evaluated, DPH seems to be slightly better than DirichletLM with respect to the IR-wise important relevance criterion (slightly better average nDCG@5 and more topics "won") while Dirichlet shows the best performance on the other quality dimensions and has a lower score variance. The TFIDF and BM25 retrieval models perform significantly worse such that args.me might have to switch its retrieval model.

(a)

Topic	Neutral Version	Biased Version
Query	plastic bottles	ban plastic bottles
Desc.	You read about the risk of plastic bottles in a newspaper article. To form a personal opinion, you search for pro and con arguments.	You watched a film describing the dangers plastic poses to nature. You want to persuade your friends to support a ban of plastic bottles, searching for suitable arguments.

(b)



(c)

Model	Rel.	Rhet.	Log.	Dial.
DirichletLM	13	**15**	**19**	**18**
DPH	**18**	**15**	11	16
TFIDF	8	5	7	3
BM25	5	5	4	6

(d)

	Rhetoric		Logic		Dialectic	
	Laymen	Expert	Laymen	Expert	Laymen	Expert
Rhetorical	–	–	0.65	**0.81**	0.60	0.75
Logical	**0.65**	**0.81**	–	–	**0.69**	**0.78**
Dialectical	0.60	0.75	**0.69**	0.78	–	–
Relevance	0.40	–	0.48	–	0.71	–

(e)

Arg. quality:		Relevance		Rhetoric		Logic		Dialectic		n = 199	
Arg. stance:		Pro	Con	Pro	Con	Pro	Con	Pro	Con	Pro	Con
Assessor Stance	Mean										
	Pro	**2.88**	2.54	**2.56**	2.35	**2.67**	2.38	**2.49**	2.21	57	48
	Con	**2.83**	2.81	2.51	**2.86**	**2.76**	2.54	**2.49**	2.43	37	35
	Neutral	2.78	**2.92**	2.54	**3.00**	2.31	**2.56**	2.23	**2.44**	13	9
	p-value										
	Pro	0.1112		0.1575		0.0880		0.0888			
	Con	0.4766		0.0908		0.1774		0.4054			
	Neutral	0.3349		0.0951		0.2612		0.2815			

(f)

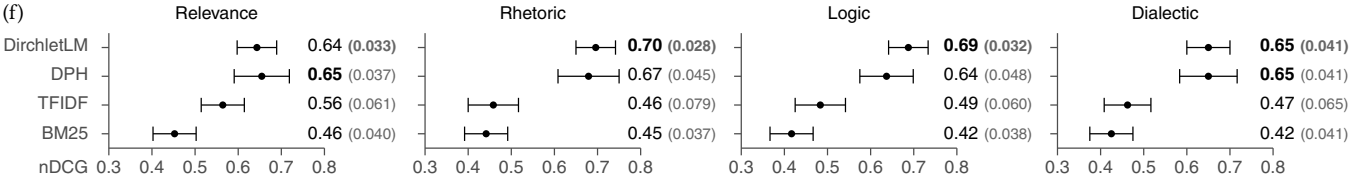


Table 1: (a) Example topic. (b) Distribution of argument assessment scores from 1 (least) to 4 (highest). (c) Amount of times a model scored highest on a topic. Total number per column can exceed 40 due to ties. (d) Pearsons' ρ between laymen and expert assessments, the latter from [21, Table 3]. (e) Mean argument scores and Mann-Whitney-test p-values cross-tabulation per argument and assessor stance. (f) Mean nDCG@5 scores along their confidence intervals and variances (in parentheses).

5 CONCLUSION AND FUTURE WORK

The web has evolved into the single most important information source for everyone who needs to make a decision, be it small or large. This places web search engines squarely at the center of many a discourse of society and demands for a fresh assessment of their role in people's decision-making processes. In this paper, we take a first step, laying the groundwork for the future development of argument search into a mature addition to retrieval technology.

Argument search raises lots of important new questions for future work: Since the notion of relevance is much more subjective, this affects search engine providers and users alike. Will providers be able to automatically judge which stances for a topic should be reflected among the top 10 result slots, and then continually defend their decision against public accusations of tampering from stakeholders not represented atop? Will users consider only that argument search engine worthy which confirms their beliefs? And if so, will providers withstand the resulting economic pressure to just give their users what they want, instead of what they need?

REFERENCES

[1] G. Albaum. 1997. The Likert scale revisited. *Int. J. of Market Research* 39(2): 1-21.
[2] G. Amati. 2006. Frequentist and Bayesian approach to information retrieval. In *Proc. of ECIR 2006*, 13-24.
[3] J. A. Blair. 2012. *Groundwork in the theory of argumentation: Selected papers of J. Anthony Blair.* Springer Science & Business Media.
[4] J. A. Blair. 2012. Rhetoric, dialectic, and logic as related to argument. *Philosophy & Rhetoric* 45(2): 148-164.
[5] A. Bondarenko, M. Völske, A. Panchenko, C. Biemann, B. Stein, and M. Hagen. 2018. Webis at TREC 2018: Common Core Track. In *Proc. of TREC 2018.*
[6] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39(2): 175-191.
[7] J. Habermas. 1984. *The theory of communicative action.* Beacon Press.
[8] I. Habernal and I. Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proc. of ACL 2016*, 1589-1599.
[9] S. Jackson. 2015. Design thinking in argumentation theory and practice. *Argumentation* 29 (3: 243-263.
[10] K. Järvelin and J. Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM TOIS* 20(4): 422-446.
[11] R. H. Johnson. 2009. Revisiting the logical/dialectical/rhetorical triumvirate. In *Proc. of OSSA 2009*, 1-13.
[12] R. H. Johnson. 2012. *Manifest rationality: A pragmatic theory of argument.* Routledge.
[13] D. Mohammed. 2016. Goals in argumentation: A proposal for the analysis and evaluation of public political arguments. *Argumentation* 30(3): 221-245.
[14] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. 2006. Terrier: A high performance and scalable information retrieval platform. In *Proc. of OSIR 2006*, 18-25.
[15] I. Persing and V. Ng. 2015. Modeling argument strength in student essays. In *Proc. of ACL 2015*, 543-552.
[16] R. Rinott, L. Dankin, C. Alzate Perez, M. M. Khapra, E. Aharoni, and N. Slonim. 2015. Show me your evidence – An automatic method for context dependent evidence detection. In *Proc. of EMNLP 2015*, 440-450.
[17] H. W. J. Rittel and M. M. Webber. 1973. Dilemmas in a general theory of planning. *Policy Sciences* 4(2): 155-169.
[18] K. Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1): 11-21.
[19] C. Stab, J. Daxenberger, C. Stahlhut, T. Miller, B. Schiller, C. Tauchmann, S. Eger, and I. Gurevych. 2018. ArgumenText: Searching for arguments in heterogeneous sources. In *Proc. of ACL 2018: Demos*, 21-25.
[20] F. H. Van Eemeren and R. Grootendorst. 2010. *Speech acts in argumentative discussions.* Walter de Gruyter.
[21] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, and B. Stein. 2017. Computational argumentation quality assessment in natural language. In *Proc. of EACL 2017*, 176-187.
[22] H. Wachsmuth, M. Potthast, K. Al Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, and B. Stein. 2017. Building an argument search engine for the web. In *Proc. of ArgMining 2017*, 49-59.
[23] H. Wachsmuth, B. Stein, and Y. Ajjour. 2017. "PageRank" for argument relevance. In *Proc. of EACL 2017*, 1117-1127.
[24] D. Walton. 2006. *Fundamentals of critical argumentation.* Cambridge Univ. Press.
[25] J. W. Wenzel. 1990. Three perspectives on argument: Rhetoric, dialectic, logic. *Perspectives on argumentation: Essays in honor of Wayne Brockriede*, 9-26.
[26] C. Zhai and J. Lafferty. 2017. A study of smoothing methods for language models applied to ad hoc information retrieval. *SIGIR Forum* 51(2): 268-276.