A Systematic Comparison of Methods for Finding Good Premises for Claims

Lorik Dumani
dumani@uni-trier.de
University of Trier
Trier, Germany

Ralf Schenkel
schenkel@uni-trier.de
University of Trier
Trier, Germany

ABSTRACT

Research on computational argumentation has recently become very popular. An argument consists of a claim that is supported or attacked by at least one premise. Its intention is the persuasion of others. An important problem in this field is retrieving good premises for a designated claim from a corpus of arguments. Given a claim, oftentimes existing approaches' first step is finding textually similar claims. In this paper we compare 196 methods systematically for determining similar claims by textual similarity, using a large corpus of (claim, premise) pairs crawled from debate portals. We also evaluate how well textual similarity of claims can predict relevance of the associated premises.

KEYWORDS

argumentation retrieval, argument search, debate portals, web arguments, similarity methods

ACM Reference Format:
Lorik Dumani and Ralf Schenkel. 2019. A Systematic Comparison of Methods for Finding Good Premises for Claims. 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/10.1145/3331184.3331282

1 INTRODUCTION

In contrast to research on argumentation, which was already been studied by Aristotle more than 2,300 years ago [4], research on computational argumentation has only recently become popular. An *argument* is, to put it simply, a *claim* or a standpoint that is supported or attacked by at least one *premise* [8], which forms a simple argument graph. Since premises can in turn be attacked or supported, often large argument networks emerge for a major claim [8]. The purpose of argumentation is the persuasion of others.

To support users arguing for or against a topic, argument search engines like ARGS[1] or ARGUMENTEXT[2] take a claim as input and return a list of premises that support or attack the query claim. These systems usually work on precomputed argument graphs that were either mined from texts or extracted from dedicated argument websites like debate.org or debatepedia.org. Premise retrieval is surprisingly difficult since frequently, a query claim and good supporting or attacking premises have only small textual overlap. Assume, for example, a user searching for premises supporting the claim "we should abandon nuclear energy". A suitable premise could be "wind and solar energy can already provide most of the energy we need", which only overlaps in the rather general term "energy". Existing methods like [10] usually use a combination of claim and premise as a retrieval unit. We argue that a more promising and principled approach than directly querying for premises is a two-stage process that first retrieves, given a query claim, matching claims from the argument collection, and then considers their premises only.

In this paper we systematically evaluate the suitability of existing term-based ranking methods for this two-stage retrieval model. Existing work [9, 10] has relied on variants of BM25 for claim retrieval, so a more systematic analysis is advisable. To build a collection of claims with associated premises, we crawled, similar to Wachsmuth et al. [10], about 60,000 arguments from four debate portals; the dataset is introduced in Section 3. In Section 4, we then perform a systematic comparison of 196 different textual similarity methods provided by Apache Lucene, using 232 query claims on the topic *energy*. Our results show that axiomatic approaches as well as DFR and many other methods significantly outperform BM25 for claim retrieval in an nDCG-based evaluation already at low cutoffs. We then examine the premise retrieval step more closely. In Section 5 we investigate in the suitability of premises of retrieved claims for the query claim, i.e., if the assumption underlying our two-stage retrieval approach is reasonable.

2 RELATED WORK

Wachsmuth et al. [10] introduce one of the first prototypes of an argument search engine called ARGS. Their system operates on arguments crawled from debate portals. Given a user query, the system retrieves, ranks, and presents premises supporting and attacking the query claim, taking similarity of the query claim with the premise, its corresponding claim, and other contextual information into account. They apply a standard BM25F ranking model implemented on top of Lucene.

Stab et al. [9] present ARGUMENTEXT, an argument retrieval system capable of retrieving topic-relevant sentential arguments from a large collection of diverse web texts for any given controversial topic. The system first retrieves relevant documents, then it identifies arguments and classifies them as "pro" or "con", and presents

[1] www.args.me
[2] www.argumentsearch.com

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331282

them ranked by relevance in a web interface. In their implementation, they make use of Elasticsearch and BM25 to retrieve the top-ranked documents.

Habernal and Gurevych [2] propose a semi-supervised model for argumentation mining of user-generated web content. In contrast to both works, we do not consider the argument mining task, but assume that we operate on a collection of arguments with claims and premises. In a followup work [3], Habernal and Gurevych address the relevance of premises. Since relevance underlies a subjective judgement, first they confronted users in a crowdsourced task with pairs of premises to decide which premise is more convincing. Then, they used a bidrectional LSTM to predict which argument is more convincing. Wachsmuth et al. [11] consider the problem of judging the relevance of arguments and provide an overview of the work on computational argumentation quality in natural language, including theories and approaches.

3 DATASET CONSTRUCTION

Since the dataset described in [10] is not publicly available, we reconstructed a similar dataset following the approach in that paper. We crawled the arguments from four debate portals, namely `debate.org`, `debatepedia.org`, `debatewise.org`, and `idebate.org`[3]. We used wget to crawl the debate portals and JSoup to process them. This resulted in overall 59,126 claims with 695,818 premises, so on average about 11.8 premises per claim. The data is available on request from the authors.

Due to the design of these portals, many collected claims are actually not statements, but questions. We have identified 50,629 of a total of 59,126 claims as questions by question marks at the end of the sentence. However, this is hardly a problem, as the questions we have observed are closed questions, i.e., they leave little room for answering. Then, Apache Lucene was used to index claims and premises of the arguments in separate indexes, using standard preprocessing without stopword removal.

Since real-life query inputs of users are rare, we drew a random sample of 232 claims that are related to the topic “energy” and used them as queries: In order to find arguments that are related to the topic energy, we trained a word-embedding-model on our corpus using DeepLearning4j. We retrieved the nearest words of the word energy, filtered out inappropriate suggestions and repeated this approach 5 times for all newly added suggestions. In the end, we obtained the following 44 words: “energy”, “nuclear”, “renewable”, “plants”, “electric”, “electricity”, “hydroelectric”, “plant”, “water”, “fukushima”, “wind”, “reactors”, “solar”, “environmentally”, “turbines”, “hydropower”, “fusion”, “chernobyl”, “uranium”, “reactor”, “carbon”, “radioactive”, “nucleus”, “atoms”, “dioxide”, “emissions”, “orangehouse”, “fossil”, “fuels”, “methane”, “fuel”, “environmental”, “trees”, “gas”, “oil”, “coal”, “environment”, “climate”, “gases”, “co2”, “emit”, “emission”, “polluting”, “warming”. This resulted in 1,529 candidate claims where at least one of these words occurred, from which we drew a random sample of 232 claims, making sure by manually inspection that they really are related to the topic energy and that duplicates were excluded. When we indexed the data, we tried to keep the corpus clean so that there were as few duplicates as possible.

Table 1: Relevance levels for claim assessment

score	meaning
5	The claims are semantically equal.
4	The claims differ in polarity, but are otherwise equal.
3	The claims differ in specificity or extent.
2	The claims address the same topic, but are unrelated.
1	The claims are unrelated.

4 EVALUATION OF CLAIM RETRIEVAL

We considered 196 different retrieval methods[4] implemented in Apache Lucene and retrieved, for each method, result claims for our 232 query claims. Note that we also tried a Doc2Vec [7] model trained on our corpus, but discarded it since its results were not satisfying in a preliminary study.

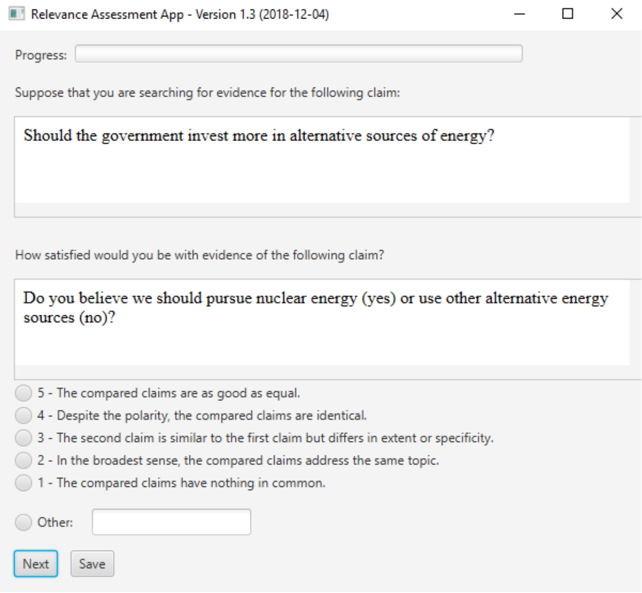
From the results, we built pools of depth 5, i.e., including any claim that appeared in the result list of any method at rank 5 or better. This resulted in 3622 (query claim, result claim) pairs, excluding pairs where the result claim was equal to the query claim. The user-perceived similarity of each (query claim, result claim) pair was independently assessed by at least two annotators on the scale from 1 to 5 (see Table 1).

In category 4 we define polarity not only as “for” or “against” but also as a neutral point of view, which is typically the case for questions. For example, the pair consisting of “*We should rely on nuclear energy*” and “*Should we rely on nuclear energy?*” would fall into category 4. Please note that the gaps between 3 and 4 as well as between 4 and 5 are much lower than the gaps between 1 and 2, as well as between 2 and 3. The underlying assumption of this scale is that all premises of claims rated 4 or 5 should apply to the query claim, whereas no premises of claims rated 1 should apply. For claims rated 3, we expect that a good number of premises match, whereas premises of claims rated 2 would only rarely match. We will verify this assumption in the following section.

The pairs were assessed by eight annotators, namely one professor, two PhD students, two Master students, and one Bachelor student in computer science, as well as one Master student and one Bachelor student in political science. The annotators were confronted with the query claim and a result claim and were asked to assess how well they expect the premises of the result claim (that were unknown to them) would match the query claim. Since we only wanted to measure the relevance of claims at this point, the actual premises were not considered at this point, but investigated later in Section 5. Figure 1 shows the application the annotators used to assess the relevance of claims. Since polarity of premises is not in the focus of this study, we collapse the levels 4 and 5 into a single level 4 for the remainder of this paper.

[3] Wachsmuth et al. also crawled `forandagainst.com`, but the website is no longer available.

[4] Apache Lucene (v. 7.6.0) provides 139 different similarity methods as well as a class called MultiSimilarity for multiple similarities. We tested all combinations of the best methods’ variants of Divergence from Randomness, Divergence from Independence, information-based models, and Axiomatic approaches as well as BM25 and Jelinek-Mercer in a first run and got $\sum_{k=2}^{6} \binom{6}{k} = 57$ new methods, resulting in 196 methods.



Relevance Assessment App - Version 1.3 (2018-12-04)

Progress:

Suppose that you are searching for evidence for the following claim:

Should the government invest more in alternative sources of energy?

How satisfied would you be with evidence of the following claim?

Do you believe we should pursue nuclear energy (yes) or use other alternative energy sources (no)?

○ 5 - The compared claims are as good as equal.

○ 4 - Despite the polarity, the compared claims are identical.

○ 3 - The second claim is similar to the first claim but differs in extent or specificity.

○ 2 - In the broadest sense, the compared claims address the same topic.

○ 1 - The compared claims have nothing in common.

○ Other:

Figure 1: Relevance assessment of claims.

Of the 7,444 individual assessments collected, 716 pairs were assessed by at least one of the annotators with a score of 4, 1,305 with a score of 3, 1,802 with a score of 2, and 3,611 with a score of 1. In order to ensure a high quality of the assessments, a third annotator assessed pairs with an absolute difference of at least 2 in the annotations; these assessments are included above. We used Krippendorff's α [6] to measure the inter-annotator agreement, yielding $\alpha = 0.803$, indicating that the assessments are reliable. As the depth of the pool of top results increased, so did the number of non-relevant results, which were almost always rated 1 or 2 by both annotators. For this reason, the agreement is relatively high for a subjective task such as relevance assessment. A typical example where two annotators disagreed (by giving ratings of 1 and 2, respectively) are the claims *Do you think that hybrid vehicles are beneficial to the environment?* and *Do you think that international trade is bad for the environment?*; here, one assessor identified the common topic 'environment', whereas the other did not see a common ground. We explain such different assessments by different expectations regarding possible premises relevant for the query claim; remember that the task was to decide whether the premises of a result claim would be relevant to the query claim.

As every pair of query claim and result claim was assessed, the final relevance value of a result claim for a query claim was computed as the mean value of the corresponding assessments.

Using the assessed pool of results as a gold standard, we evaluated the performance of the 196 retrieval methods under consideration for the claim retrieval task, using nDCG@k [5] with cutoff values $k \in \{1, 2, 5\}$ as quality metric. To compute nDCG, the ratings from the ground truth were renormalized by subtracting 1, yielding an interval of [0, 3] for relevance grades. Table 2 shows an excerpt of the results of this experiment, focusing on the best methods for each cutoff and the two baseline methods BM25 and Language Models with Jelinek-Mercer smoothing. The individual ranks of the various methods are in brackets behind the values. In this table, combined similarities for n different similarities are denoted by $\mathrm{MS}_{method_1,method_2,\ldots,method_n}$. The results clearly show that the BM25 scoring method used in previous works is usually not an ideal choice. This is especially true for cutoff 5, which is a realistic cutoff for a system that aims at finding the top-10 premises. Here, the best similarity method is a MultiSimilarity consisting of F1LOG (Axiomatic approaches for IR) and DFR (divergence from randomness)[5], which clearly improves over both BM25 and language models. Only the improvement over language models is statistically significant (by a two-sided paired t-test) with p <2.6E-6=0.05/19110 if one corrects for multiple comparisons with the conservative Bonferroni correction [1]; the p-value for the comparison to BM25 is 5E-5.

Table 2: $nDCG@\{1, 2, 5\}$ **for claim retrieval.**

Retrieval Function	*nDCG@1*	*nDCG@2*	*nDCG@5*
$\mathrm{MS}_{F1LOG,DFRI(ne)BZ(0.3)}$	0.7515 (18)	0.7486 (18)	**0.8355** (1)
$\mathrm{MS}_{F1LOG,BM25(k1=1.2,b=0.75)}$	0.7410 (37)	**0.7608**(1)	0.8171 (18)
DFR GB2	**0.7635** (1)	0.7525 (8)	0.8117 (36)
DFR I(ne)BZ(0.3)	0.7439 (26)	0.7571 (2)	0.8274 (3)
F1LOG	0.7428 (33)	0.7451 (26)	0.8278 (2)
⋮	⋮	⋮	⋮
BM25(k1=1.2,b=0.75)	0.7451 (24)	0.7303 (101)	0.7944 (135)
⋮	⋮	⋮	⋮
LM Jelinek-Mercer($\alpha = 0.2$)	0.7206 (115)	0.7210 (137)	0.7881 (161)

5 EVALUATION OF PREMISE RELEVANCE

We now focus on the second step of the two-stage retrieval framework, retrieving the premises of claims similar to the query claim. Our goal here is to verify the common hypothesis that claims highly similar to the query claim also have premises that are highly relevant for the query claim; this hypothesis is not only the foundation of our proposed two-stage ranking approach, but was made, for example, also by Wachsmuth et al. [10].

To systematically approach this question, we formed triples of the form (query claim, result claim, result premise) from the pool constructed in the previous sections, where the result premise is a premise of the result claim. We grouped the triples according to the relevance of the result claim to the query claim, forming groups of the relevance ranges $[n, n + 0.5)$ for $n \in \{1, 1.5, 2, 2.5, 3, 3.5\}$ and [4, 4], which yielded seven groups. Then, we randomly drew 100 (query claim, result claim, result premise) triples from each group and had two annotators manually assess the relevance of the result premise for the query claim (without seeing the result claim), resulting in 1,400 assessments.

Annotators could choose between either *not relevant* or *relevant* with three different stances: *query with neutral stance*, *premise with same stance as query* and *premise with opposite stance as query*. We considered a premise to be relevant as far as it is related to the statement in the broadest sense. The inter-annotator-agreement of this classification procedure, measured with Krippendorff's α,

[5] The parameters for DFR in the first place are the following: Inverse expected document frequency [mixture of Poisson and IDF], ratio of two Bernoulli processes, term frequency normalization provided by a Zipfian relation.

Table 3: Premise relevance relative to claim relevance.

relevance interval of claims	relevant premises (in %)	discordant ratings (in %)	non-relevant premises (in %)	other (unclear, spam,...)
[4, 4]	**79.00**	3.00	13.00	5.00
[3.5, 4.0)	**71.00**	5.00	23.00	1.00
[3.0, 3.5)	**61.00**	6.00	30.00	3.00
[2.5, 3.0)	**53.00**	4.00	41.00	2.00
[2.0, 2.5)	17.00	7.00	**74.00**	2.00
[1.5, 2.0)	2.00	6.00	**92.00**	0.00
[1.0, 1.5)	6.00	3.00	**91.00**	0.00

was 0.9. It thus seems that this assessment task was “easier” in the sense that agreement was easier to achieve.

As we did with claims before, we ignore the stances of premises since we only want to focus on their relevance, and many claims of our dataset do not have a stance anyway. We thus consider only binary relevance for premises from now on.

Table 3 shows the results of our study, including the percentage of premises where the annotators agreed on relevance or non-relevance, but also the cases where they did not agree. In general, these results support the observation that the more relevant a claim for the query is, the more relevant premises it yields. So if a search engine performs well at the claim retrieval task, it should also perform well at the subsequent premise retrieval task; the initial hypothesis is thus validated. However, it is interesting to see that even for claims that were assessed before as perfectly relevant for the query claim (rating 4), only 79% of their premises are considered relevant for the query claim. At the same time, even non-relevant claims can sometimes yield relevant premises, which is also surprising. We examined these cases and observed that the non-relevant premises of highly relevant claims either get off the point and are not even relevant to the original claim, or the premises are very specific to the original claim. An example for this is a premise for the query claim *Is adaptation a priority requiring greater focus/funding?* taken from the result claim *Should governments focus on adaptation to global warming over mitigation?* (with rating 4); here, the premise gets off the point because the user explains that he does not believe in global warming at all.

The relevant premises for claims with little relevance originate from the general applicability of some premises. An example is the query claim *Do you agree that natural gas can be our emancipation from foreign oil?* and the result claim *Should the process of fracturing rock to obtain oil and natural gas be banned in the United States?* (with rating in [1.0, 1.5)) where the premise describes that the process of fracturing would be safe, which could be seen as supporting premise of the query claim as well.

6 CONCLUSION AND FUTURE WORK

Retrieving good premises for claims is an important, but difficult problem for which no good solutions exist yet. This paper has provided some insights that a two-stage retrieval process that first retrieves claims, and then ranks their premises can be a step towards a solution. The best premises are found for the most similar claims, according to assessments by human annotators, is already good. Although we already come to a large amount of statements in the corpus with the choice of the topic energy, it is nevertheless necessary to add other topics with completely different contexts such as education or medicine in order to determine to what extent the results can be generalized.

The methods from Apache Lucene were used as a starting point. We will consider other methods, especially from the field of machine learning. Our future work will include the second step of our two-stage retrieval model, that is the clustering and ranking methods for premises, which may include textual similarity of the premise with the query claim, but also premise popularity across claims. We will also examine additional quality-based premise features [11] such as convincingness or correctness. We provide a public Web application as an interface to our system at argumentsearcher.uni-trier.de.

ACKNOWLEDGMENTS

We would like to thank Manuel Biertz, Alex Witry, Mirko Lenz, Anna Ludwig, Premtim Sahitaj and Stefan Ollinger for their help in the annotations. We would also like to thank Markus Nilles for implementing the web application and Ralph Bergmann for his valuable input. This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft - DFG) (Grant No. 375342983).

REFERENCES

[1] Norbert Fuhr. 2017. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. SIGIR Forum 51, 3 (2017), 32–41. https://doi.org/10.1145/3190580.3190586

[2] Ivan Habernal and Iryna Gurevych. 2015. Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse. In Proc. 2015 Conf. Empirical Methods in Natural Language Processing (EMNLP). 2127–2137. http://aclweb.org/anthology/D/D15/D15-1255.pdf

[3] Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In Proc. 54th Ann. Meeting of the Assoc. for Computational Linguistics (ACL). http://aclweb.org/anthology/P/P16/P16-1150.pdf

[4] Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational Argumentation Meets Serious Games. In Proc. 2017 Conf. on Empirical Methods in Natural Language Processing (EMNLP). 7–12. https://aclanthology.info/papers/D17-2002/d17-2002

[5] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20, 4 (2002), 422–446. https://doi.org/10.1145/582415.582418

[6] Klaus Krippendorff. 1970. Estimating the reliability, systematic error, and random error of interval data. Vol. 30. 61–70 pages. Issue 1.

[7] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In Proc. 31th Int. Conf. on Machine Learning, (ICML). 1188–1196. http://jmlr.org/proceedings/papers/v32/le14.html

[8] Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. IJCINI 7, 1 (2013), 1–31. https://doi.org/10.4018/jcini.2013010101

[9] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for Arguments in Heterogeneous Sources. In Proc. 2018 Conf. North American Chapter of the Assoc. for Computational Linguistics (NAACL-HTL). 21–25. https://aclanthology.info/papers/N18-5005/n18-5005

[10] Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an Argument Search Engine for the Web. In Proc. 4th Workshop on Argument Mining (ArgMining@EMNLP). 49–59. https://aclanthology.info/papers/W17-5106/w17-5106

[11] Henning Wachsmuth, Benno Stein, Graeme Hirst, Vinodkumar Prabhakaran, Yonatan Bilu, Yufang Hou, Nona Naderi, and Tim Alberdingk Thijm. 2017. Computational Argumentation Quality Assessment in Natural Language. In Proc. 15th Conf. European Chapter of the Association for Computational Linguistics (EACL). 176–187. https://aclanthology.info/papers/E17-1017/e17-1017