University of Amsterdam at CLEF 2020

Notebook for the Touché Lab on Argument Retrieval at CLEF 2020

Mahsa S. Shahshahani and Jaap Kamps

University of Amsterdam
{m.shahshahani,kamps}@uva.nl

Abstract This paper documents the University of Amsterdam's participation in CLEF 2020 Touché Track. This is the first year this track has been introduced at CLEF, and we were attracted to participate in it due to its potentialities for Parliamentary debates we are currently working on. This track consists of two tasks: Conversational Argument Retrieval and Comparative Argument Retrieval. We submitted a run to both tasks. For the first task, we used a combination of the traditional BM25 model and learning to rank models. BM25 model helps to retrieve relevant arguments, and learning to rank model helps to re-rank the list and put *strong*er arguments on top of the list. For the second task, Comparative Argument Retrieval, we proposed a pipeline to re-rank documents retrieved from Clueweb using three features: PageRank scores, web domains, and argumentativeness. Preliminary results on 5 queries have shown that this heuristic pipeline may help to achieve a balance among three important dimensions: relevance, trustworthiness, and argumentativeness.

1 Introduction

We believe that we passed the era in which search engines were supposed to only give us a ranked list of documents or answers and they have more potentialities to help us in decision making process. Argument retrieval task has been defined to formulate this problem. Touché track at CLEF [3] offers an opportunity to work on this interesting problem having access to a debate corpus from two different points of view: Looking for different views about a problem in debates between opponents and supporters of a controversial issue, and looking for comparative opinions about different alternative. This track consists of a different task (two tasks in total) for each point of view and we submitted a run to both tasks.

In this paper, we cover both tasks; first, we give a high-level summary of the first task, followed by our detailed approach. Then, we cover the same for the second task. Finally, we will conclude the paper by mentioning our main contributions and findings.

We will add results section whenever the results would be out.

2 Conversational Argument Retrieval

For detailed information about CLEF track's experimental setup, we refer to the overview paper [4] and to the track homepage.[1] However, we provide a high-level summary to make this paper self-contained.

2.1 Task definition

The goal of this task is to retrieve relevant arguments from online debate portals, given a query on a controversial topic.

Corpus Args.me [1] has been created by crawling arguments from 7 debate websites. It includes 387,606 arguments taken from 59,637 debates. A search engine based on Elasticsearch has been set to make it easier to work with this corpus [10]. This search engine ranks arguments using BM25 ranking algorithm. Different approaches can be used later to re-rank these retrieved arguments.

Queries 50 controversial topics have been picked for this task. Each topic has both pro and con arguments in the corpus.

Quality Assessment Proposed approaches are supposed to retrieve "strong" arguments. An argument is consiagstuhl ered *strong* if it is topically relevant, logically cogent, rhetorically well-written, and useful to help in stance-building process. Here, we define these assessment dimensions taken from [9] and the annotation guidelines for Dagstuhl-15512 ArgQuality Corpus.[2]

Topical relevance: As every other ranking task, retrieved arguments should provide the user with relevant information about the query.

Besides relevance, in general, there are three main dimensions for assessing the quality of arguments: logic, rhetoric, and dialectic.

Logical cogency: An argument with acceptable premises that are relevant and sufficient to the argument's conclusion is considered "cogent" [7].

Rhetorical well-writtenness: An argument is called "rhetorically well-written" if it is effective and successful in persuading a target audience of a conclusion [2].

Dialectic: An argument is considered reasonable if it contributes in the users' stance-building process regards a given issue in a way that is acceptable to everyone.

2.2 Our Approach

We treated this task as a re-ranking problem. In the absence of training data, we have to use unsupervised approaches. However, we used an existing debate dataset created for studying argument quality assessment to train a classifier. First, we describe this corpus. Later, we explain our approach in three consecutive steps.

[1] https://events.webis.de/touche-20/shared-task-1.html

[2] http://www.arguana.com

Corpus Dagstuhl-15512 ArgQuality [9] includes 20 arguments for each of 16 queries. Three annotators have annotated these 320 arguments on 15 dimensions with three labels. However, we only use four dimensions: cogency, effectiveness, reasonableness, and overall quality. The set of labels includes ordinal scores from 1 (low) to 3 (high) for all dimensions. We used majority voting technique to get the label for each dimension, and substituted the ones labeled as 3 with 2 as there are a very few samples with label "3" in the corpus .

Approach We created the final ranked list in three steps. Before explaining each step, we explain the way we represented arguments.

Argument Representation: We used pre-trained BERT-base [6] model from HuggingFace Transformers framework in python to represent arguments. As BERT model imposes a limit on the length of documents after tokenization, we used the first 512 tokens in an argument if its length exceeds this limit.

First step- Ranking: BM25 is the traditional unsupervised ranking model which scores the relevancy of documents (here arguments) in regards to queries based on the frequency of common terms between the query and argument. We ranked arguments for each topic based on BM25 using args.me search engine.

Second step- Classification: We trained a classifier on Dagstuhl-15512 ArgQuality corpus to recognize and label cogency, well-writtenness, reasonableness and overall quality of each argument. Later, we applied this classifier to all retrieved arguments in the ranked list from the fist step.

We got the majority voting for each dimension; in all of the arguments the majority for all four dimensions are the same. It is aligned with the conclusion in the main paper [9] which indicates that cogency, effectiveness, and reasonableness correlate strongly with overall quality, and also much with each other.

We trained two classifiers on 90% of data: A decision tree classifier, and an SVM classifier. The accuracy on the remaining 10% of data has been shown in Table 1. We used scikit-learn framework for python 3 to train both classifiers. For decision tree, we used "gini" criterion, and set minimum required samples to split to 2. For SVM, we used "rbf" kernel, and set regularization parameter to 1. We selected SVM classifier for the further step.

Third step-Re-ranking: In the third step, we re-ranked retrieved arguments from the first step using learning-to-rank models. We use the output of step 2 as a feature in step 3 to re-rank the arguments. We trained three different learning to rank models: Ranknet, RandomForests, and LambdaRank. We used RankLib[3] library with default sets of parameters to apply these models. In order to train learning to rank models, we used argument representations based on BERT model, the output of the second step, and two additional features based on named entities. We defined two binary features indicating the presence of numerical named entities (percent, quantity, money) and other

[3] https://github.com/codelibs/ranklib

entities (person, location, organization) in the argument. We showed the number of arguments with and without these entities in Figure 1 and Figure 2. These figures show the difference in the distribution of each of these features in strong (label=2) and weak (label=1) arguments. Arguments using these kinds of entities are more likely providing users with persuasive and effective information to make their stance, and this can lead to a more probability to be labeled as "strong".

We trained learning to rank models on Dagstuhl dataset, and applied the best model (Ranknet) to re-rank retrieved arguments from args.me dataset for all 50 topics in the shared task. We trained models on 90% of data and reported accuracy on the remaining 10% of data in Table 2.

Table 1. classifiers

Model	Accuracy
Decision Tree	0.43
SVM	0.53

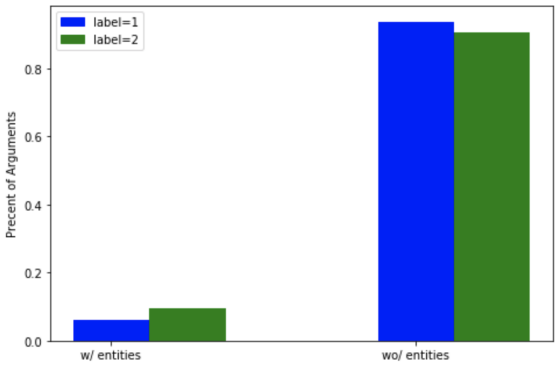

Figure 1. Having/lacking numerical named entities (percent, money, quantity)

Table 2. LTR

Model	NDCG@1	NDCG@5
RankNet	**0.873**	**0.9587**
Random Forests	0.7333	0.9155
LambdaRank	0.7778	0.9291

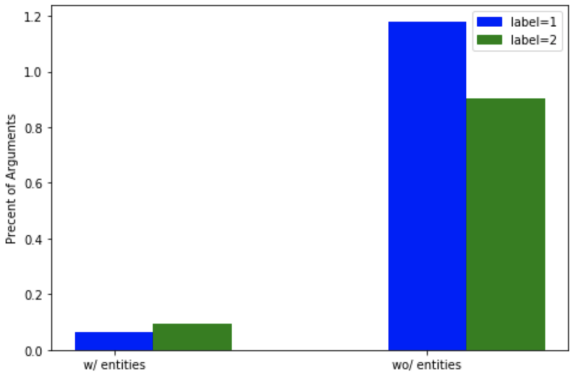

Figure 2. Having/lacking other named entities (person, organization)

2.3 Results

In the final relevance judgments, 30 documents for each topic have been annotated using 6 labels: -2,1,2,3,4 and 5.

Classifier We ran our classifier that was trained on Dagstuhl-15512 ArgQuality corpus on arguments from judgments to see if the results are correlated with the final judgment. Unfortunately, we observed that the classifier does not work well and return label '1' for more than 90% of the judged arguments. This suggests that the classifier does not play a role in the final results of our learning to rank model. This is not surprising as the dataset we trained our SVM classifier on is very different from the test set. The arguments in the training set are very short and consist of only one to three sentences, while we are labeling a set of complete documents in the test set. However, as the BERT representation has a limit on the size of the input text, for the longer texts we only used the first 512 tokens of each argument. Using more advanced approaches such as averaging over sliding windows of 512 tokens might make the classifier useful.

Entities Similar to Figure 1 and Figure 2, we looked into the distribution of numerical and other types of entities in relevant and non-relevant documents. We considered documents with label '-2' as non-relevant, and other documents in the judgment file as relevant. Results have been shown in Figure 3 and 4. As it is obvious from the figures, the distribution of entities is the opposite of what we observed in the dataset we used at the time of developing our model. However, it still can be used as a feature as it shows a little difference between relevant and non-relevant documents.

Query Length We looked into per query results to get an insight into the way our model works. We used NDCG@5 metric as it has been the main metric for the shared task.

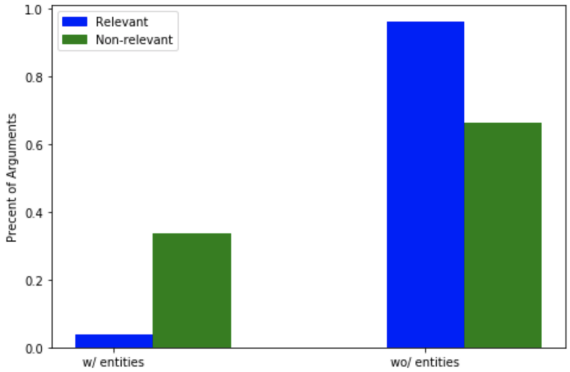

Figure 3. Having/lacking numerical named entities (percent, money, quantity)

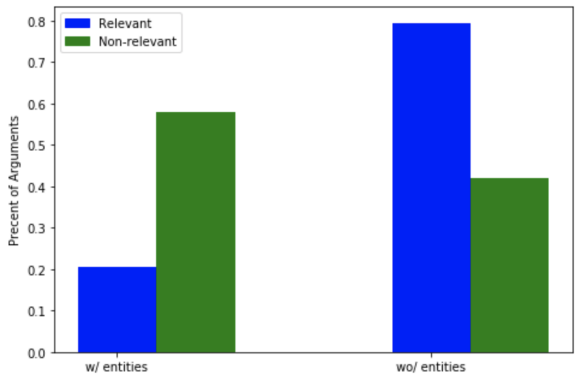

Figure 4. Having/lacking numerical named entities (person, organization)

Table 3. Results-NDCG@10 metric

Model	**NDCG@1**	**NDCG@5**	**NDCG@10**	**MAP**
UvATask1LTR	0.5214	0.5548	0.3709	0.1129

As we used the whole topic (without removing stop words), the model works better for shorter queries (Figure 5).

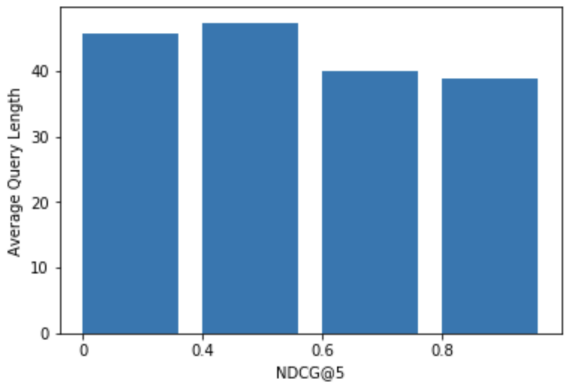

Figure 5. Average Query Length for the different ranges of NDCG@5)

3 Comparative Argument Retrieval

Similar to the previous section, for detailed information about this task's experimental setup, we refer to the overview paper [4] and to the track homepage.[4] However, we provide a high-level summary.

3.1 Task definition

The goal of this task is to retrieve and rank documents from web that help to answer a comparative question from "everyday life".

Corpus Clueweb12 is a dataset created by crawling 733,019,372 web documents seeded with 2,820,500 urls from Clueweb09 [5]. We used a publicly available search engine [8] based on Elasticsearch to retrieve documents from Clueweb12 based on BM25 ranking model.

Queries 50 comparative topics from everyday life have been picked for this task.

3.2 Our approach

We treated this task as a re-ranking problem. In the absence of training data, we have to use unsupervised approaches. We labeled retrieved documents for 5 topics to have an insight of how our heuristic approach works.

In the first step, we used ChatNoir search engine to rank documents retrieved for each topic. In the second step, we used three different features to re-rank them. Here, we introduce the features we used, followed by our heuristic approach to combine them and create the final ranked list.

[4] https://events.webis.de/touche-20/shared-task-2.html

Argumentativeness We trained a simple SVM classifier based on data from args.me corpus and Clueweb to distinguish between argumentative and non-argumentative documents. Similar to the first task, we used BERT-based model from HuggingFace Transformers library for Python to represent documents, and we used the first 512 tokens in a document if its length exceeds the limit of BERT model.

To train the classifier, we used a small sample from each corpus. These samples are created by submitting all 50 controversial queries in the first task to both corpora and got up to 100 documents for each query. Then, we manually removed argumentative documents from the sample taken from Clueweb and considered the remaining documents as negative examples. All retrieved documents from args.me corpus have been considered as positive examples. The final training set consists of 3000 positive and 3000 negative examples. Then, We trained a simple SVM classifier on 80% of the data, and evaluated it on the remaining 20% of documents. It achieved 87% in terms of accuracy. All parameters for the argumentativeness classifier have been set to their default values in Scikit-learn[5] library for Python.

Web domains Clueweb has been formed by crawling web documents with some post-filters. But, the goal of this task is to retrieve documents including personal opinions or suggestions. Thus, documents from particular domains like Wikipedia are not desirable. On the contrary, documents from discussion forums, debate websites, and blogs can be very helpful. Having this intuition in mind, we defined a binary feature that indicates if the source URL for a discussion contains 'forum' or 'blog' terms to give a bonus to web pages from discussion forums or blogs.

PageRank Although desired documents are those from discussion forums and personal blogs, they should also be trustworthy. To take trustworthiness into account, we used page rank scores to prioritize documents taken from more reliable sources.

In ChatNoir search engine, every returned document has been associated with a PageRank score. We directly used these returned scores.

Re-ranking We introduced three features, and our goal is to re-rank documents based on a combination of these features (argumentativeness, domain addresses , and PageRank scores).

To generate the final ranked list, we make a heuristic ranking pipeline in four steps:

- **The 1st step:** The initial ranked list is taken from ChatNoir search engine. ChatNoir retrieves documents from Clueweb and ranks them using traditional BM25 ranking model. We use the whole topic title as the query. We also examined submitting the query after removing stop words or just using the entities in the topic title. But, using the whole topic title worked better.
- **The 2nd step:** Page-rank scores are used to re-rank the list from the first step in descending order. This may result in putting a document, initially ranked very low, on top of the list. In order to avoid this, moving documents in the ranked list is limited to a maximum of 10 positions.

[5] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

Table 4. Preliminary Results- NDCG@10 metric

Model	Relevance	Argumentativeness	Trustworthiness
initial	0.87	0.71	0.81
pagerank	0.89	0.66	0.87
domain	0.84	0.72	0.80
argumentative classifier	0.80	0.84	0.79
mixed	0.84	0.78	0.82

- **The 3rd step:** Web domain are used to re-rank ranked list from the second step. To do this, the documents with positive domain feature (which means the document is taken from a blog or discussion forum) are put on top of the list. This has been performed within every 10 documents in the list. We split the list into chunks of 10 documents and we keep their relative positions.
- **The 4th step:**All retrieved documents in the ranked list from the third step are classified using the argumentativeness classifier we have trained. Documents classified as *positive* are put on top of the list. We keep their relative positions. This has been performed within every chunk of 10 documents in the list.

We put this limits on moving documents in the list with this intuition in mind that relevance should be prioritized in comparison with trustworthiness and argumentativeness.

3.3 Results

Since the final judgment file does not consist of separate judgment lists for three different dimensions (relevance, trustworthiness, and argumentativeness), we include our preliminary results too.

Preliminary Results: To gain an insight into the effectiveness of our heuristic model, we manually labeled 10 retrieved documents for 5 queries. We labeled documents using three labels: 0 for non-relevant, non-argumentative, or untrustworthy; 1 for relevant, argumentative, or trustworthy, and 2 for highly-relevant, highly-argumentative, or highly trustworthy.

We evaluated the top 10 documents for each ranked list: BM25, re-ranked by PageRank scores, re-ranked by web domains, re-ranked by argumentativeness, and mixed model.

Evaluation results have been reported in Table 3. The heuristic mixed model does not achieve the same performance as BM25 in terms of relevance, the same performance as the argumentative classifier model in terms of argumentativeness, and the same performance as the PageRank model in terms of trustworthiness. But, it seems that it struck a balance among all three dimensions.

Table 5. Final Results

Model	NDCG@1	NDCG@5	NDCG@10
initial	0.5068	0.4480	0.4196
pagerank	0.4723	0.4256	0.4087
domain	0.4100	0.3845	0.4132
argumentative classifier	0.4281	0.4123	0.4023
UvATask2SVM	0.5000	0.4464	0.4185

Final Results: The final results for every step of our model have been reported in Figure 5. Since three different aspects had been defined for evaluation of this task, We expected to receive three judgment sets. However, the judgment file issued only the overall relevance, which we sacrificed for the other two dimensions.

Initially, we processed the data and implemented our models based on the three different aspects which had been defined previously. Those include relevance, trustworthiness, and argumentativeness. Given the fact that we received the overall relevance file only, we extracted our results purely based on the relevance factor for the evaluation of this task. This categorically implies that the obtained results could have been more optimized if we had performed our preliminary assessment based on one aspect only.

4 Conclusion

This paper documents our first participation in the Touché 2020 Track. We explained our approaches for both tasks in the track: Conversational Argument Retrieval, and Comparative Argument Retrieval.

For the Conversational Argument Retrieval task, we used an existing argument quality assessment dataset to train a classifier and re-rank arguments based on the output of this classifier. We showed that named entities are important features to distinguish between *strong* and *weak* arguments on the preliminary data. But, the final results show that neither classifier nor entities do not help. However, it is worth mentioning that the classifier had been trained on a completely different set of arguments. So, if we train it on the same data from args.me, it might be useful for the final results. This needs further investigations to be proved.

For the Comparative Argument Retrieval task, we introduced three features to re-rank arguments taken from Clueweb. We proposed a pipeline to combine different aspects (relevance, trustworthiness, and argumentativeness) to create the final ranked list. Preliminary results have shown that this heuristic pipeline may successfully strike a balance between all three dimensions. However, the final evaluation has been done only on relevance. This caused our method to be sub-optimal.

We hope and expect that the valuable bench-marking data created at Touché track will be of great value to motivate, and greatly facilitate, further research into argument retrieval.

Acknowledgments

This research was supported in part by the Netherlands Organization for Scientific Research (NWO, grant # CISC.CC.016, ACCESS project). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

1. Ajjour, Y., Wachsmuth, H., Kiesel, J., Potthast, M., Hagen, M., Stein, B.: Data acquisition for argument search: The args.me corpus. In: Benzmüller, C., Stuckenschmidt, H. (eds.) KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11793, pp. 48–59. Springer (2019). https://doi.org/10.1007/978-3-030-30179-8_4, https://doi.org/10.1007/978-3-030-30179-8_4
2. Blair, J.A.: Groundwork in the theory of argumentation: Selected papers of J. Anthony Blair, vol. 21. Springer Science & Business Media (2012)
3. Bondarenko, A., Fröbe, M., Beloucif, M., Gienapp, L., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2020: Argument Retrieval. In: Working Notes Papers of the CLEF 2020 Evaluation Labs (Sep 2020)
4. Bondarenko, A., Hagen, M., Potthast, M., Wachsmuth, H., Beloucif, M., Biemann, C., Panchenko, A., Stein, B.: Touché: First shared task on argument retrieval. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12036, pp. 517–523. Springer (2020). https://doi.org/10.1007/978-3-030-45442-5_67, https://doi.org/10.1007/978-3-030-45442-5_67
5. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 web track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009. NIST Special Publication, vol. 500-278. National Institute of Standards and Technology (NIST) (2009), http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423, https://doi.org/10.18653/v1/n19-1423
7. Johnson, R.H., Blair, J.A.: Logical self-defense. International Debate Education Association (2006)
8. Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., Welsch, C.: Chatnoir: a search engine for the clueweb09 corpus. In: Hersh, W.R., Callan, J., Maarek, Y., Sanderson, M. (eds.) The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012. p. 1004. ACM (2012). https://doi.org/10.1145/2348283.2348429, https://doi.org/10.1145/2348283.2348429
9. Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: Computational argumentation quality assessment in natural language. In: Proceedings

of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers. pp. 176–187 (2017). https://doi.org/10.18653/v1/e17-1017, https://doi.org/10.18653/v1/e17-1017

10. Wachsmuth, H., Potthast, M., Khatib, K.A., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., Stein, B.: Building an argument search engine for the web. In: Habernal, I., Gurevych, I., Ashley, K.D., Cardie, C., Green, N., Litman, D.J., Petasis, G., Reed, C., Slonim, N., Walker, V.R. (eds.) Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017. pp. 49–59. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/w17-5106, https://doi.org/10.18653/v1/w17-5106