Retrieving Comparative Arguments using Ensemble Methods and Neural Information Retrieval

Notebook for the Touche Lab on Argument Retrieval at CLEF 2021

Viktoriia Chekalina[1,2], Alexander Panchenko[1]

[1]Skolkovo Institute of Science and Technology, Moscow, Russian Federation
[2]Philips Innovation Lab Rus, Moscow, Russian Federation

Abstract

In this paper, we present a submission to the Touché lab's Task 2 on Argument Retrieval for Comparative Questions [1, 2]. Our team Katana supplies several approaches based on decision tree ensembles algorithms to rank comparative documents in accordance with their relevance and argumentative support. We use PyTerrier [3] library to apply ensembles models to a ranking problem, considering statistical text features and features based on comparative structures. We also employ large contextualized language modelling techniques, such as BERT [4], to solve the proposed ranking task. To merge this technique with ranking modelling, we leverage neural ranking library OpenNIR [5].

Our systems substantially outperforming the proposed baseline and scored first in relevance and second in quality according to the official metrics of the competition (for measure NDCG@5 score). Presented models could help to improve the performance of processing comparative queries in information retrieval and dialogue systems.

Keywords

comparative argument retrieval, natural language processing, neural information retrieval

1. Introduction

On a daily basis, people face the problem of choosing between two entities - which phone is more reliable, which juice contains less sugar, which hotel is better for a holiday. Domain-specific comparison systems, like WolframAlpha or Diffen, solve this problem partly and rely on structured data, which limits the number of cases it can be used.

On the other hand, the Web contains a vast number of opinions and objective arguments that can facilitate the comparative decision-making process. It creates the need of developing an open-domain general system that could process such information. The issue is to retrieve from a set of documents relevant, supportive and credible arguments. The aim of the proposed work is to retrieve from ClueWeb12 [1] corpus documents and re-rank them, considering argumentation for or against one option or the other.

The contribution of our work is the following: we are first to use ensemble methods based on mixed statistical and comparative features to the document ranking; we are first to use neural information retrieval approach to the task of argument retrieval; we propose a model

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

[1]http://lemurproject.org/clueweb12

outperforming the baseline and yielding the first and the second-best result according to the relevance and quality metric, respectively.

2. Related work

The most relevant to this work is the previous shared task Touche 2020 [6]. 17 participants took part in the competition and submitted 41 runs. Various approaches were tested by these participants, including methods based on extraction of structures corresponding to claims and premises, assessing argument quality, representation of documents by language models, expansion of the query by similar words. The ranking function from search engine ChatNoir [7] based on BM25F [8] approach was used as a baseline.

Only a few of the submitted solutions can slightly improve the baseline. The best overall approach in the previous competition was the method based on query extension and reranking documents by relevance, credibility, and supportive quality [9].

This work is based on our run submitted in the previous version of the Touche shared task [10]. In this work, we used a pre-trained language model to find relevance between the query and document. Extraction of comparative structures and counting the number of comparative sentences in a document help us to assess the quality of relevant arguments.

Therefore, the problem of argument retrieval arises in other scenarios. Comparative argumentation machine CAM [11] retrieves comparative sentences with respect to accepted objects and comparison aspects. The paper [12] explores the influence of context on an argument detecting system and proves the performance increasing related to it.

3. Datasets and experimental design

3.1. Datasets

The organizers provided 50 comparative questions (topics), for which we should obtain documents containing convincing arguments for or against one or another option. Topics for the competition are available online. [2]

In addition, 50 topics and corresponding relevance annotations of the previous year's competition [13] were given for supervised learning. These documents were also retrieved from ChatNoir and ranked manually to 0 (not relevant), 1 (relevant) or 2 (highly relevant) scores. We use this data to train and set up models based on the decision trees and fine-tune the BERT ranker. Besides, last year's teams submissions were available too.

Unfortunately, this data is not insufficient for fitting large supervised ranking models, for example, based on the BERT technique. In this case we use adjacent question-answering dataset called Antique [14]. This dataset consists of the questions and answers of Yahoo! Webscope L6 and contains 2,626 open-domain non-factoid questions and 34,011 manual relevance annotations.

The example of query and ranked answers are in Table 6, Table 7 in Appendix A. It might be noticed that Antique dataset has a different set of ranking scores - 0, 1, 2 instead of 1, 2, 3, 4 - so we rewrite Antique ranks in accordance with the following mapping 1→0, 2→1, 3→1, 4→2.

[2] https://webis.de/events/touche-21/shared-task-2.html

3.2. Evaluation setup

We use every topic as a query in ChatNoir [7] search engine and extract up to 100 unique documents from the ClueWeb12 corpus. We clean documents' bodies from HTML tags and markups and ranked them using one of the developed approaches described below.

As auxiliary data, the organizers provided the topics of the previous year's competition. For each proposed topic, a set of documents from ChatNoir was retrieved and labelled as described above. We use this data to train developing models and valid composing approaches. In the validation phase, we split the ranked data into 40 topics in train and 10 in validation.

In the run phase, we execute produced solutions on web evaluation platform Tira [15]. In this stage to fit the model we use ranked data from the previous year entirely and predict rank for current proposed topics. The runs were evaluated using the NDCG metrics based on the human judgements of the submitted runs. Retrieved documents were judged in accordance with two criteria: (i) document relevance, (ii) whether sufficient argumentative support is provided [16].

4. Document ranking using ensembles of trees

In this section, we use ensembles of trees as a supervised machine learning technique to solve ranking problems. We choose either pointwise regression tree algorithms, like Random Forest, or boosted tree algorithms like XGBoost and LightGBM. In the cases of LightGBM model we employ LambdaMART [17] objective. It combines cost function derived from minimizing the number of inversions in ranking (LambdaRank [18]) and objective for building gradient boosted decision trees (MART [19]). We use PyTerrier platform for information retrieval.[3] It simplifies the extraction of the text features and allows expressing retrieval experiments [20].

4.1. Feature extraction

For our ranking ML methods, we use features that came from 3 origins described below: (i) ranking features extracted by PyTerrier, (ii) specific comparative features, (iii) score from ChatNoir system based on custom BM25 scoring function.[4]

4.1.1. Features extracted by PyTerrier

PyTerrier provides measure of matching query-document texts by several models. Among these models are statistical measures (TF-IDF), mesures based on language models (Heimstra, Diriclet), measures based on occurrence of a document depending on the fields that the term occurs in (BM25F, PLF). The list of all possible models are available at the cite [5]. Among these varieties we have chosen BM25, Heimstra, DFIC, DPH, TF-IDF, DiricletLM, PL2 for our exploration.

We applied each of the selected methods sequentially and independently to the training set, ranked documents by the obtained scores and evaluated the ranking on the validation set. The result of these tests is in Table 1. We have chosen 3 methods with the most promising results, and these 3 methods combine 3 features.

[3] https://pyterrier.readthedocs.io/en/latest/index.html
[4] https://www.elastic.co/guide/en/elasticsearch/reference/current/index-modules-similarity.html
[5] http://terrier.org/docs/current/javadoc/org/terrier/matching/models/package-summary.html

Table 1
Results on validation set for text features in PyTerrier models.

Method	**BM25**	Heimstra	**DFIC**	DPH	**TF-IDF**	DiricletLM	PL2
NDCG@5	**0.3637**	0.3616	**0.3642**	0.3110	**0.3637**	0.3307	3703

4.1.2. Comparative features

We focus not only on finding high relevant documents as on finding documents with a comparison of one object relative to another. The work [21] assumes that the comparative issue can be represented by comparative structures - objects for comparison, comparative aspects and predicates. We take the sequence-labelling model suggested in the cited paper and applied it to the query. It helps us to define objects for comparison for every topic. Then we apply the model to document and get a comparative feature set.

The feature `is_retrieved` describes are there any comparative structures in the document at all. Characteristic `objs_score` defines how many objects from query are found in document (0, 1 or 2). Feature `asp_pred_score` is counted in the following way: if at least one object from a query is in the document, every word in the document labelled as an aspect or predicate increases the score to 0.5. Finally, we combined defined features with scores obtained from the ChatNoir system, and a resulting feature vector for pair query-document is {`score_pl2`, `score_tf`, `score_bm`, `score_dfic`, `baseline_scores`, `is_retrieved`, `ap_score`, `objs_score`}.

4.2. Models

4.2.1. Random Forest

We use the Random Forest model imported from Sklearn and wrapped by the PyTerrier pipeline. To find the best setup, we vary the number of estimators from 10 to 150, the value 20 gives the best valid score NDCG@5 of 0.408.

4.2.2. XGBoost

We also wrapped gradient boosting library from Sklearn to PyTerrier class and tune hyperparameters by setting the learning rate from $1e{-}4$ to 0.1 and max_depth from 4 to 16. The best setup is learning rate 0.01, max_depth 6 and gives NDCG@5 0.547.

4.2.3. LightGBM

In the case of LightGBM, we vary the number of leaves from 8 to 20 and the learning rate from 0.001 to 0.1. The best configuration with num_leaves equal to 15 and a learning rate equal to 0.1 gives 0.579 score.

The feature importance of the resulting model is in Table 2. It can be seen that the most significant feature is the score retrieved from the ChatNoir, then there is a Divergence from Independence based on Chi-square [22] and the existence of comparison objects in the document.

Table 2
Feature importance in the proposed LightGBM model

Feature	Pl2	TF-IDF	BM25	Dfic	ChatNoir	has comp	objs_score	asp_pred
Importance	1.76	1.19	1.51	2.3	20.8	0	1.66	1.51

5. Document ranking using neural information retrieval based on BERT

Contextualized language models such as BERT can be much more efficient for ranking tasks because they contain vast relationships between language units. In the proposed work we use a reranking model from OpenNIR [5][6] based on "Vanilla" Transformer architecture [23].

5.1. Text representation

BERT receives a query and document and processes it jointly. A distinctive feature of the BERT reranker is injection token similarity matrices on each layer, which considerably improves performance [24].

5.2. Training process

First, we pretrain this reranker on the Antique dataset. We clean this dataset from incorrect symbols and makeups. We also left from the dataset documents of length more than 300 characters, since the length of the ChatNoir retrieves usually does not exceed 300. The training process lasted for 500 epochs with 0.001 learning rate and 56 objects in every batch. Finally, our model gives NDCG@5 0.3362 on a validation set. We fine-tune the model on 40 train topics from the previous year for 50 epochs with the same configuration. Fine-tuning increased the score on validation up to 0.412.

6. Results

6.1. Results on validation set

The result for every proposed approach obtained on the validation part of data from the previous year competition is in Table 3. We also evaluate the previous year's baseline on the validation set. The best scores come from the LightGBM model, which also outperforms the baseline. XGBoost has fewer scores, Random Forest as a simple algorithm has the smallest score. Bert overtakes Random Forest a little.

In the right column, we also added the time required to train each model. It can be seen that the ensemble-based models have approximately the same time complexity, while the Bert requires much more time to train.

[6] https://github.com/Georgetown-IR-Lab/OpenNIR

Table 3
Results on validation set.

Method	NDCG@5	Time, ms
Random Forest	0.408	127.168
XGBoost	0.547	128.848
LightGBM	**0.572**	131.244
Bert Ranker	0.412	1560.947
Baseline'20	0.534	-

6.2. Results on test set

For final testing, the retrieved documents were labelled manually with a score from 0 to 3. Judgment was carried out in two independent criteria: the relevance of the document to the given topic and the quality of the text. Quality criterion includes good language styling, easy reading and proper sentence structure, the absence of typos and alliteration.

For each criterion, a separate file with the assessor's scores is available. The results of two evaluations are presented in the Table 4 and Table 5. The runs of our team Katana have the best result between all teams in terms of relevance and the second result in terms of the text quality.

As in the validation set, XGBoost and LightGBM give the best performance. It is well explained, since the loss of these models based on the ranking quality functions, NDCG in the XGBoost case and LambdaMART in the LightGBM case. The first model describes relevance a bit better (0.489) and has first place among the whole participant. For quality, conversely, LightGBM is better. It archives 0.684 and takes second place in a quality table, slightly surrendering to Top 1. The random forest method has scores just below the baseline in both cases. It can be explained by a more elementary algorithm for building an ensemble. Bert gives a quite good result for quality and weak for relevance. Perhaps the data from the adjacent task (factoid QA) used for the training is the reason for not a very accurate solution.

Table 4
NDCG@5 scores on runs for relevance for Katana team, baseline and Top-2 approach

Method	NDCG@5
Random Forest	0.393
XGBoost (Top 1)	**0.489**
LightGBM	0.460
Bert Ranker	0.091
ChatNoir baseline	0.422
Thor team (Top 2)	0.478

Table 5
NDCG@5 scores on runs for quality for Katana team, baseline and Top-1 approach

Method	NDCG@5
Random Forest	0.630
XGBoost	0.675
LightGBM (Top 2)	**0.684**
Bert Ranker	0.466
ChatNoir baseline	0.636
Rayla team (Top 1)	0.688

7. Conclusion

In this paper, we present our solution to the Argument retrieval shared task. We pay attention to ensembles methods and use statistic approaches, language modelling and comparative structure extraction to retrieve features for it. We also use a neural reranker based on the Bert technique to use information from a contextualized model in our task.

The best results were obtained by gradient boosting methods, training on ranking cost function: XGBoost and LightGBM. The proposed approaches outperform baseline and take first and second places in relevance and quality ranking, respectively. Bert contextualized model shows the need for large learning data.

Acknowledgments

This work is partially supported by the project "ACQuA: Answering Comparative Questions with Arguments" (grants BI 1544/7-1 and HA 5851/2-1) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999). We thank Maik Frobe for providing the support of the software runs in the TIRA system.

Appendix A: Examples of training data

Table 6
Example of query and document with different relevance in Touche task dataset

Query	Document	Rank
What is better for the environment, a real or a fake Christmas tree?	Disease and condition content is reviewed by our medical review board real or artificial? There is so much confusing information out there about which is better for your health and the environment.	2
	You may think you're saving a tree, but the plastic alternative has problems too. Which is "greener" an artificial Christmas tree or a real one?	1
	This entry is part 25 of 103 in the series eco-friendly friday november 28th's tip christmas trees: stuck between choosing a real Christmas tree or a fake one?	0

Table 7
Example of query and document with different relevance in Antique dataset

Query	Document	Rank
Why do we put the letter k on the words knife and knob, knee?	They are saxon words. Knife would have been pronounced ker-niff.	4
	As a guess I would say that historically "kn" would have been pronounced differently to "n" and that time has altered the way the words are pronounced.	3
	Because English is a funny language.	2
	I don't really (k)now!	1

Appendix B: Examples of ranking results

In this appendix you can find examples Top-3 ranked documents in accordance to LightGBM and Baseline approaches.

Table 8
Example of documents with the different relevance to query "Is admission rate in Stanford higher than that of MIT?"

Is admission rate in Stanford higher than that of MIT?	
LightGBM Top-3	Baseline Top-3
1. Stanford and Harvard have a similar admissions rate of about 7%. MIT comes with a somewhat greater rate of success admitting just under 10% or 1742 for the class of 2015. Harvard, Stanford and MIT are global leaders in culture, commerce and governmental policies.	1. Stanford and Harvard have a similar admissions rate of about 7%. MIT comes with a somewhat greater rate of success admitting just under 10% or 1742 for the class of 2015. Harvard, Stanford and MIT are global leaders in culture, commerce and governmental policies
2. For more than a decade, i have served as an admissions officer for MIT. In that time, i've read more than 10,000 applications and have watched thousands of new students enter MIT. It is a privilege to work at the most dynamic and exciting university in the world.	2. For more than a decade, i have served as an admissions officer for MIT. In that time, i've read more than 10,000 applications and have watched thousands of new students enter MIT. It is a privilege to work at the most dynamic and exciting university in the world.
3. Our primary enhancement was targeted at families earning less than $75,000 — making mit tuition free and eliminating	3. All of this factual information, plus a lot of other detail, can be found in the mit admissions literature. In fact, this year, mit will award $74 million in undergraduate aid.

Table 9
Example documents with the different relevance to query "Which smartphone has a better battery life: Xperia or iPhone?"

Which smartphone has a better battery life: Xperia or iPhone?	
LightGBM Top-3	Baseline Top-3
1. 1. The power saver app that will turn down settings when battery life is low to get as much juice out of the battery as possible. Sony has set the benchmark with its 12 megapixel camera inside the Xperia S.	1. The iPhone 4 is apple's thinnest smartphone yet, but offers a much better screen, faster processor, video calling, and many other enhancements.
2. How to increase the battery life of apple's iPhone 4s many of those with an iphone 4s have complaints about the battery life. Apple has acknowledged these problems, and is working to fix them.	2. Sony Xperia's review: an above average smartphone gizmotraker', as far as battery life is concerned, it last about 7 hr 30 min in talktime, 450 hrs in standby.
3. Sony Ericsson includes an 8gb card in the sales package the Sony Ericsson Xperia arc s has below average battery life. Most users will get around 24 hours of life out of the Xperia. X27's 1600mah battery before it needs a recharge, but heavy users may need an injection of power before then.	3. How to increase the battery life of Apple's Iphone 4s many of those with an iphone 4s have complaints about the battery life. Apple has acknowledged these problems, and is working to fix them.

References

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

[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] S. MacAvaney, C. Macdonald, N. Tonellotto, IR from Bag-of-words to BERT and Beyond through Practical Experiments: An ECIR 2021 tutorial with PyTerrier and OpenNIR, in: Proceedings of the 43rd European Conference on Information Retrieval Research, 2021, pp. 728–730.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[5] S. MacAvaney, OpenNIR: A Complete Neural Ad-Hoc Ranking Pipeline, in: J. Caverlee, X. B. Hu, M. Lalmas, W. Wang (Eds.), WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020, ACM, 2020, pp. 845–848. URL: https://doi.org/10.1145/3336191.3371864. doi:10.1145/3336191.3371864.

[6] 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, 2020, pp. 384–395. doi:10.1007/978-3-030-58219-7_26.

[7] M. Potthast, M. Hagen, B. Stein, J. Graßegger, M. Michel, M. Tippmann, C. Welsch, ChatNoir: A Search Engine for the ClueWeb09 Corpus, in: B. Hersh, J. Callan, Y. Maarek, M. Sanderson (Eds.), 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2012), ACM, 2012, p. 1004. doi:10.1145/2348283.2348429.

[8] S. E. Robertson, H. Zaragoza, M. J. Taylor, Simple BM25 extension to multiple weighted fields, in: D. A. Grossman, L. Gravano, C. Zhai, O. Herzog, D. A. Evans (Eds.), Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004, ACM, 2004, pp. 42–49. URL: https://doi.org/10.1145/1031171.1031181. doi:10.1145/1031171.1031181.

[9] T. Abye, T. Sager, A. J. Triebel, An open-domain web search engine for answering comparative questions, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

URL: http://ceur-ws.org/Vol-2696/paper_130.pdf.

[10] V. Chekalina, A. Panchenko, Retrieving comparative arguments using deep pre-trained language models and NLU, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_210.pdf.

[11] M. Schildwächter, A. Bondarenko, J. Zenker, M. Hagen, C. Biemann, A. Panchenko, Answering comparative questions: Better than ten-blue-links?, in: L. Azzopardi, M. Halvey, I. Ruthven, H. Joho, V. Murdock, P. Qvarfordt (Eds.), Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10-14, 2019, ACM, 2019, pp. 361–365. URL: https://doi.org/10.1145/3295750.3298916. doi:10.1145/3295750.3298916.

[12] M. Fromm, E. Faerman, T. Seidl, TACAM: topic and context aware argument mining (2019) 99–106. URL: https://doi.org/10.1145/3350546.3352506. doi:10.1145/3350546.3352506.

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

[14] H. Hashemi, M. Aliannejadi, H. Zamani, W. B. Croft, ANTIQUE: A non-factoid question answering benchmark 12036 (2020) 166–173. URL: https://doi.org/10.1007/978-3-030-45442-5_21. doi:10.1007/978-3-030-45442-5_21.

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

[16] L. Braunstain, O. Kurland, D. Carmel, I. Szpektor, A. Shtok, Supporting human answers for advice-seeking questions in CQA sites, in: N. Ferro, F. Crestani, M. Moens, J. Mothe, F. Silvestri, G. M. D. Nunzio, C. Hauff, G. Silvello (Eds.), Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings, volume 9626 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 129–141. URL: https://doi.org/10.1007/978-3-319-30671-1_10. doi:10.1007/978-3-319-30671-1_10.

[17] Q. Wu, C. J. Burges, K. M. Svore, J. Gao, Adapting bboosting for information retrieval measures, Information Retrieval 13 (2010) 254–270. URL: https://www.microsoft.com/en-us/research/publication/adapting-boosting-for-information-retrieval-measures/.

[18] C. Burges, R. Ragno, Q. Le, Learning to Rank with Nonsmooth Cost Functions, in: B. Schölkopf, J. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems, volume 19, MIT Press, 2007. URL: https://proceedings.neurips.cc/paper/2006/file/af44c4c56f385c43f2529f9b1b018f6a-Paper.pdf.

[19] J. Friedman, Stochastic Gradient Boosting, Computational Statistics & Data Analysis 38 (2002) 367–378. doi:10.1016/S0167-9473(01)00065-2.

[20] C. Macdonald, N. Tonellotto, Declarative Experimentation in Information Retrieval using PyTerrier, in: K. Balog, V. Setty, C. Lioma, Y. Liu, M. Zhang, K. Berberich (Eds.), ICTIR '20:

The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020, ACM, 2020, pp. 161–168. URL: https://dl.acm.org/doi/10.1145/3409256.3409829.

[21] V. Chekalina, A. Bondarenko, C. Biemann, M. Beloucif, V. Logacheva, A. Panchenko, Which is better for deep learning: Python or MATLAB? Answering Comparative Questions in Natural Language, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 302–311. URL: https://www.aclweb.org/anthology/2021.eacl-demos.36.

[22] I. Kocabas, B. Dincer, B. Karaoğlan, A Nonparametric Term Weighting Method for Information Retrieval Based on Measuring the Divergence from Independence, Information Retrieval (2013) 1–24. doi:10.1007/s10791-013-9225-4.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[24] S. MacAvaney, A. Yates, A. Cohan, N. Goharian, CEDR: Contextualized Embeddings for Document Ranking (2019) 1101–1104. URL: https://doi.org/10.1145/3331184.3331317. doi:10.1145/3331184.3331317.