Retrieving Comparative Arguments using Deep Language Models

Notebook for the Touché Lab on Argument Retrieval at CLEF 2022

Viktoriia Chekalina[1], Alexander Panchenko[1]

[1]Skolkovo Institute of Science and Technology

Abstract

In this paper, we present a submission to the Touché lab's Task 2 on Argument Retrieval for Comparative Questions. Our team Katana employs approaches based on pre-trained deep language model architecture ColBERT [1]. This BERT-based architecture is adapted to the text ranking task by learning to represent both queries and documents as vectors and measuring the similarity between them. We use a model trained on a question-answering dataset MSMARCO, with the proposed weights and weights pre-trained by us. We also customize ColBERT for the comparative retrieval domain by fine-tuning the model on the data from the previous years' Touché competitions. The proposed experiments verify the usefulness of the transfer learning from a large pre-trained ranking language models to the problem of arguments extraction for comparative topics. Ours solutions rank third in both relevance, quality, and stance prediction evaluations.

Keywords

comparative argument retrieval, natural language processing, neural information retrieval

1. Introduction

In everyday life, people are constantly faced with the task of comparing two options: which of the phone models is more reliable, which fuel is environment-friendly, which drug is the most effective. The decision-making process is based not only to comparing the structural features of objects, as suggested, for example, by WolframAlpha[1] or Diffen[2], but on considering people's opinions. The problem of searching on the web for documents with argumentative support for compared objects is a subset of information retrieval tasks problem.

The Touché lab's Task 2 on Argument Retrieval in 2022 [2] proposes to select passages from a corpus of 1 million texts that are most relevant to the user's comparative queries, as well as to determine their position - which object in the text is proposed as the most suitable. We employ neural-network based approach with a simplified scheme for comparing query and document embeddings. In addition to using the pre-trained large language model, we further trained the model on documents ranked for comparative queries.

On the validation dataset, the approach shows competitive performance, but less than the ensemble-based method from the previous year [3]. This work shows the possibility and

CLEF'22: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

[1] https://www.wolframalpha.com

[2] https://www.diffen.com

efficiency of the neural network technique based on the matching of the query and document representations relatively to a specific comparative case of informational retrieval.

2. Related work

The most relevant to this work are the previous shared tasks Touche 2020 [4] and Touche 2021 [5]. These tasks aimed to rerank documents, retrieved by ChatNoir [6] System as candidates to a comparative request answers. Multiple teams submitted their runs to the shared task as presented in the technical reports of CLEF.[3]

The main difficulty in finding relevant documents on the web is the large size of the text corpus. Traditionally, search engine systems depict documents using statistic-based features, the computation of which is not complex.

For example, the baseline of the 2021 and 2020 comparative shared tasks is created on the BM25F [7] – a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each passage. This baseline performed well - only a few teams in previous years' competition could overcome it. One of the best improvements solutions was the decision tree ensembles over statistics and comparison features [3], deploying in PyTerrier [8] library.

A large volume of texts imposes a limitation on the use of neural networks for ranking documents in a corpus. There are two ways of neural approaches to information retrieval tasks: representation-based models [9] and interaction-based models [10]. The first one computes the representation of the topic and passage separately and only counts the score of interaction for the pair. Interaction-based methods match the query and document in a token or phrase-level. This set of methods is more expensive but most effective. In the proposed paper we deploy architecture, which combines the advantages of both these methods.

3. Data and experimental design

3.1. Data provided for the task

The organizers offer the participants 50 comparative questions (topics), for which it was necessary to extract and rank passages from the text corpus. Topics for the competition are available online[4]. The organizers also provide a corpus of about 0.9 million texts for passage extraction. For stance detection, every topic comprises objects that are compared in it. For stance detection support, a dataset created from comparative questions of the MSMARCO dataset[5] is proposed. The dataset includes relevant answers with highlighted objects of comparison in it and their position in the documents. Every text in a dataset has a detected stance.

For model validation purposes, the task presents 100 topics and corresponding relevance annotations of the previous year's competition [4, 5]. These documents were also retrieved from ChatNoir and ranked manually to 0 (not relevant), 1 (relevant), or 2 (highly relevant) scores. The 2020 year assessment contains a common ranking, last year's competition has a separate

[3] http://ceur-ws.org/Vol-2696, http://ceur-ws.org/Vol-2936
[4] https://webis.de/events/touche-22/shared-task-2.html
[5] https://microsoft.github.io/msmarco

judgment for relevance and quality. We use this data to fine-tune the model to comparative sub-task in document retrieval. Besides, last year's team submissions are available too.

3.2. Datasets

The standard learning object for argument ranking consists of a triple: query, positive passage (relevant text), negative passage (irrelevant text). Reading comprehension dataset MS-MARCO (Microsoft Machine Reading Comprehension) [11] includes 1,010,916 anonymized questions from Bing's query and 8 million passages extracted from the search system Bing. For the training BERT-based model we use MSMARCO-Passage-Ranking, which comprises triplets from the mentioned questions and passages.

We use data from the previous years' Touche tasks to generate a validation dataset and dataset for fine-tuning the ColBERT model. For every topic, we retrieve up to 100 texts from the ClueWeb12 [6] corpus using the ChatNoir [6] system, according to Tocuhe'20-21 task rules. The validation dataset was created on 10 topics from 2021 with corresponding quality and relevance qrels. The rest 40 topics and 50 topics from 2020 produce data for adapting the pre-trained model for text ranking in terms of argumentative objects comparison.

The 2020 year task topics have only one assessment dimension in qrels. If the score in this is 1 or 2, we treat this text as relevant. Irrelevant pairs were selected from documents with ratings less than 1 or from the search results for different topics, provided that they were not presented in the search results for the current query. In the case of an assessment of 21 years, there are separate judgments among two axes: quality and relevance. We calculate a sum of a quality and relevance score and consider relevant documents having a score equal to or more than 3, otherwise - irrelevant. The statistic of mentioned datasets is in Table 1.

Table 1
Statistics of datasets used in training from scratch and fine-tuning.

Dataset	Task	Number of triples
MSMARCO-Passage-Ranking	train	39 780 810
Dataset based on Touché 2021	fine-tune	46 450

3.3. Evaluation setup

For document ranking, we use ColBERT [1] model, pre-trained in several ways. Using the model, in the test stage, we created an index of all documents in the provided collection of text passages. Using this index, we select the top-k most relevant texts to each of the topics. We use auxiliary information about objects under comparison to find them in every ranked document and define document stance using Comparative argumentation machine CAM [12] functionality. We execute produced solutions on the web evaluation platform Tira [13]. The retrieved documents will be assessed manually for both metrics: general relevance and comparison quality. Relevance depicts proximity to the topic and the presence of sufficient argumentative support. Quality refers to good structuring, understandable news, and text styling.

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

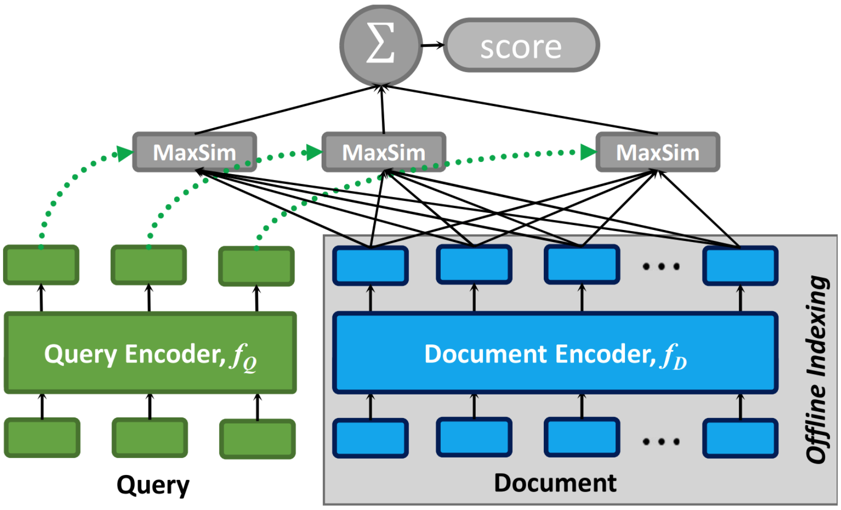

Figure 1: The scheme of Late Interaction matching is used in ColBERT architecture. The similarity of query and document is the sum of the scores between every query token and the most similar document token. Source of the image: [1].

In the validation phase, we use topics of the previous year's competition as queries. The corpus on which the model builds the index consists of documents from the Chat Noir issue that are relevant to topics. We retrieve documents for every question and compare them to official qrels judgments.

4. Document ranking

4.1. Document ranking with Late Interaction over BERT representations

The main architecture we used in the retrieving document task is Contextualized Late Interaction over BERT (ColBERT). ColBERT provides a trade-off between representation-based models with low computational cost and well-performed token interaction-based models. Actually, for approaches with a full interaction matrix between query and document tokens, ColBERT reduces complexity by affording a convolution over the documents' token.

The query and document processing in ColBERT architecture contains 2 steps:

- To encode query, we add $[Q]$ after $[CLS]$ token, process padded query by BERT, apply convolution and normalization
- To encode document, we add $[D]$ after $[CLS]$ token, process padded passage by BERT, apply convolution and normalization, also filter out punctuation symbols and other tokens unimportant under retrieval task.
- The conception of Late Interaction (Fig. 1) from the entire document considers only the token that has the maximum similarity with the given query token. The document relevance is estimated as a sum of maximum similarities across all query tokens.

- For retrieving in a large-scale set of passages, the faiss library [14] for the efficient similarity search is used.

Thus, the ColBERT approach fine-tunes BERT main encoder and learns from the scratch linear layers, filter and embeddings for $[q]$ and $[d]$ symbols. Leveraging on triplets of query, document with high relevance and document with low relevance $< q, d^+, d^- >$, the model optimizes the pairwise softmax cross-entropy loss.

4.2. ColBERT models

For passage retrieval in the Touche task, we use three different types of pre-trained ColBERT architecture.

ColBERT original The first is a checkpoint, generated at the University of Glasgow [7] on MSMARCO triples using instruction from the official ColBERT repository [8].

ColBERT from scratch We also pre-trained ColBERT architecture, provided in repository, from scratch by ourselves. We use L2 distance between a query and document instead of cosine similarity, since the original paper noted that the faiss index works faster on a square distance. The training process was carried out in a 3 epochs with the learning rate $3\mathrm{e}{-6}$, batch size 64, passage length no more than 180, query length 32, similarity $l2$, and took about two weeks on a single GPU card.

CoBERT fine-tune We also tried to fine-tune the resulting model on data for a comparative question-answer system obtained from information from past competitions and described in section 1. The pre-training procedure was carried out with the following parameters: learning rate $1\mathrm{e}{-7}$, batch size 64, passage length no more than 180, query length 32, similarity $L2$. The weights are updated using the AdamW optimizer during 10 epochs.

5. Stance detection

An additional challenge within the task was to determine the stances of retrieved documents. Stance defines the document's attitude towards the compared objects: pro first object, pro-second object, neutral, or the absence of attitude. To detect the stance of a given document, we note objects from topic auxiliary data, found them in the document, and consider text between objects' locations. Comparative Argumentative Machine (CAM) offers the possibility of classification those pieces of text. It decodes them into feature vectors using Infersent [15] and applies a pre-trained XGBoost classifier to features [12]. The output of CAM is considered to be a document stance class.

[7] http://www.dcs.gla.ac.uk/~craigm/colbert.dnn.zip

[8] https://github.com/stanford-futuredata/ColBERT

6. Results

We run the proposed approaches in two stages: in the validation stage the model retrieves and ranks documents for the previous year's topic over the ChatNoir output, and in the test stage the model ranks passages for a given topics over proposed corpus, at the same time designating their stance.

6.1. Results on validation set

Table 2
NDCG@5 results for quality and relevance of retrieved document on validation set.

Method	Quality	Relevance
Baseline'21	0.427	0.649
Best Answer'21	0.421	0.591
ColBERT original	0.413	0.474
ColBERT from scratch	0.342	0.314
ColBERT fine-tune	0.322	0.365

The result for every proposed approach obtained on the validation part of data from the previous year's competition is in Table 2. We compare ColBERT-based approaches to the previous year's baseline and LGBM Ranker, considered the best answer. The best scores come from the frequency-based feature baseline approach, the second place belongs to the ensembles over statistic and comparative features set. Pre-trained ColBERT provides results slightly worse in terms of quality. In terms of accuracy, the decrease is more significant, but the same in order as the difference between the first and second places scores. ColBERT, trained by our team from scratch, provides a worse result than pre-trained ColBERT. Fine-tuning this version on the dataset from the previous year's task gives a noticeable increase in relevance, but makes the model perform slightly worse on quality. This may be due to the properties of the Touche-based dataset used for model fine-tuning. It contains passages, less complete and grammatically correct than MSMARCO objects, but at the same time they are more suitable specifically for the comparative subset of questions.

6.2. Results on test set

The retrieved documents were assessed manually for two dimentions. The first criteria is relevance - how opportune and supportive answer is contained in passage, the second is rhetorical quality - good styling and well understoodness of the text. The results also contains the F1 macro clssification scores for the stance detection. The results for three criteria for our tean Katana and Top-1 approch in each metrics are in Tables 3.

For the ranking document task, ColBERT, trained on the MSMARCO dataset has the best performance according to fine-tuning the model. The difference between the model with downloaded weights and the model trained by us from scratch is not significant. Pre-trained model achive 3rd place in terms of relevance, while model trained from scratch has 3rd place in the quality table. Fine-tuning comparative data impairs the results. It may be due to the quality difference between texts from the main and fine-tuning data - in the MSMARCO case, well-formed natural language passages were composed by humans on the basis of the search system outputs. [11]. The quality of the stance detection towards the objects expectedly depends on the ranking performance - the ColBERT with pre-trained weights also takes third place.

Table 3
Final evaluation scores on the test set for Katana team as compared to the Top-1 approaches.

Method	NDCG@5 relevance	NDCG@5 quality	F1 stance detection
ColBERT original	0.618 (Top-3)	0.643	0.229 (Top-3)
ColBERT from scratch	0.601	0.644 (Top-3)	0.221
ColBERT fine-tune	0.574	0.637	0.212
Top-1 approach	0.758	0.774	0.313

7. Conclusion

We present our solution for Argument Retrieval for Comparative Questions – a ranking task over a corpus of textual passages. In our submission, we use large pre-trained neural models which match representations of an input text document and a query. More specifically, we experiment with the ColBERT model, based on computational effective late interaction architecture. We employ a model, pre-trained on the question answering dataset MSMARCO. For adapting the model to a particular comparative case, we fine-tune it on a dataset built on a ranked document from previous years' competition. We also detect stances of every ranked text by using the classification functionality of the comparative argumentative machine: it determines the polarity of text between two objects by the pre-trained InferSent model.

According to the manual assessment, the best quality in all metrics - both ranking and stance detection – comes from a model trained on the large dataset MSMARCO showing that the pre-trained model already allows to answer comparative questions decently turning out to be a strong baseline. A straighforward procedure of fine-tuning of this model with the comparative questions worsens the ultimate quality of the model.

Source code of our experiments is available online.[9]

Acknowledgments

We thank Maik Frobe for providing the support of the software runs in the TIRA system.

[9] https://github.com/sayankotor/touche

References

[1] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over BERT, CoRR abs/2004.12832 (2020). URL: https://arxiv.org/abs/2004.12832. arXiv:2004.12832.

[2] A. Bondarenko, M. Fröbe, J. Kiesel, S. Syed, T. Gurcke, M. Beloucif, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2022: Argument Retrieval, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2022, p. to appear.

[3] V. Chekalina, A. Panchenko, Retrieving comparative arguments using ensemble methods and BERT, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 2354–2365. URL: http://ceur-ws.org/Vol-2936/paper-211.pdf.

[4] A. Bondarenko, M. Fröbe, M. Beloucif, L. Gienapp, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2020: Argument Retrieval, 2020, pp. 384–395. doi:10.1007/978-3-030-58219-7_26.

[5] 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 - extended abstract, in: D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II, volume 12657 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 574–582. URL: https://doi.org/10.1007/978-3-030-72240-1_67. doi:10.1007/978-3-030-72240-1_67.

[6] 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.

[7] 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.

[8] 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.

[9] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, L. Heck, Learning deep structured semantic models for web search using clickthrough data, in: Proceedings of the 22nd ACM International Conference on Information Knowledge Management, CIKM '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 2333–2338. URL: https://doi.org/10.1145/2505515.2505665. doi:10.1145/2505515.2505665.

[10] B. Mitra, F. Diaz, N. Craswell, Learning to match using local and distributed representations of text for web search, in: R. Barrett, R. Cummings, E. Agichtein, E. Gabrilovich (Eds.), Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017, ACM, 2017, pp. 1291–1299. URL: https://doi.org/10.1145/3038912.3052579. doi:10.1145/3038912.3052579.

[11] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated machine reading comprehension dataset, CoRR abs/1611.09268 (2016). URL: http://arxiv.org/abs/1611.09268. arXiv:1611.09268.

[12] 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.

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

[14] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2019) 535–547.

[15] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised learning of universal sentence representations from natural language inference data, CoRR abs/1705.02364 (2017). URL: http://arxiv.org/abs/1705.02364. arXiv:1705.02364.