Retrieving Comparative Arguments using Deep Language Models

Viktoriia Chekalina, Alexander Panchenko

Skolkovo Institute of Science and Technology

9 aug 2022

Task

- Given a set of 50 topics with a comparative query, i.e. question like "What is better X or Y ?"
- For each topic:
  - Retrieve relevant passages in a corpus of about 0.9 million texts.
  - For every retrieved passage, detect stance towards the compared objects X and Y.
- Two criteria for evaluating retrieved passages:
  - Relevance - how sensible and supportive the answer is.
  - Quality - good styling and well understoodness of the text.

Approach to ranking

ColBERT - the ranking model which fits deep language model BERT for informational retrieval. ColBERT provides *Late Interaction* concept:

- the query and the document are encoded separately
- resulting similarity of pair is a sum of similarity of every query token and the closest document token

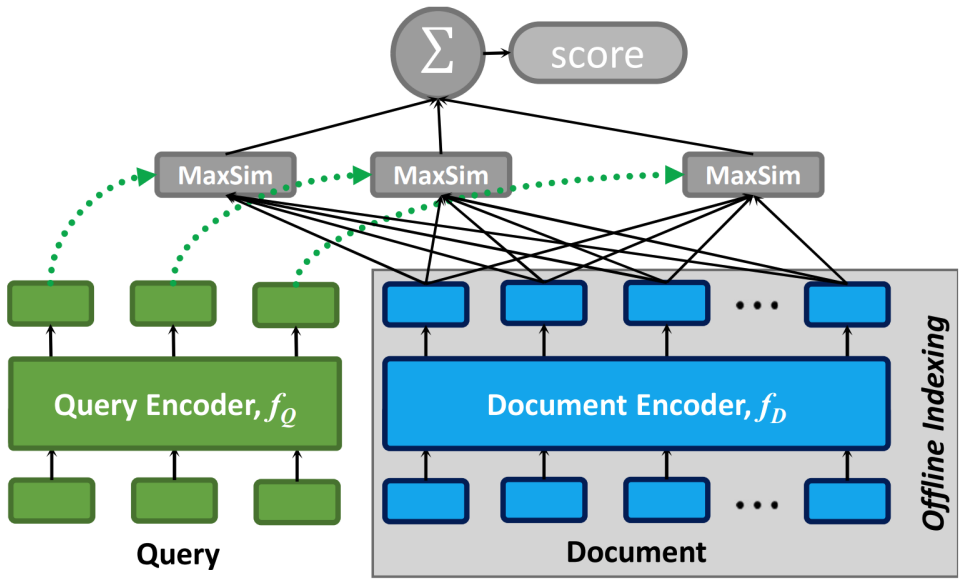

Figure: The scheme of Late Interaction matching is used in Colbert architecture. Source of the image: `https://github.com/stanford-futuredata/ColBERT`.

Approach to ranking

We use three different types of pre-trained ColBERT:

- **ColBERT original** - checkpoint, generated at the University of Glasgow [1] on MSMARCO dataset
- **ColBERT from scratch** - checkpoint, pre-trained by us on MSMARCO dataset
- **ColBERT fine-tuned** - checkpoint, pre-trained on MSMARCO, fine-tuned then on the previous year Touche data

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

Datasets and Stance Detection

Datasets used in training from scratch and fine-tuning.

- MSMARCO (Microsoft Machine Reading Comprehension) is based on queries and passages from the Bing system
- In the fine-tune dataset, the triples $< q, d^+, d^- >$ consist of queries from previous years, texts assessed by high score, texts set by low score or text related to other queries.

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

Stance Detection

To detect passage polarity, we consider text between two object entrances and classify it using Comparative Argumentative Machine (CAM) [2].

[2] `https://ltdemos.informatik.uni-hamburg.de/cam`

Results on Validation and Test sets

Table: 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

Table: 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

Conclusion

- We apply BERT-based architecture ColBERT to the comparative text ranking task.
- We apply the XGBoost classifier for the stance detection of retrieved documents.
- The best scores, which place third places in the leaderboard, gives the model with weights pre-trained on the MSMARCO dataset.