Retrieving Comparative Arguments using Ensemble Methods and Neural Information Retrieval

Viktoriia Chekalina, Alexander Panchenko

Skolkovo Institute of Science and Technology, Philips Innovations Lab RUS

24 sent 2021

Task

- Given a set of topics with comparative query, i.e. question like “What is better X or Y ?”
- For each topic:
  - retrieve documents from ClueWeb12[1] corpus by ChatNoir[2] search engine.
  - rank documents in accordance to most full and reasonable comparison.

Table: Example of query and documents with different relevances

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

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

[2] https://www.chatnoir.eu/doc

Evaluation & Approaches

Evaluation setup

- **Test set**: 50 topics with comparative questions
- Organizers are also provide 50 topics with corresponding relevance annotations of the previous year's competition. We split it to:
  - **Train set**: 40 topics
  - **Valid set**: 10 topics

Approaches to ranking

- Ensembles of trees
- Reranking Bert model

Ensembles of trees

Information retrieval (IR)

PyTerrier[3] platform for information retrieval:

- Extraction of the text features
- IR adaptation of ensembles model
- Expressing IR experiments

Ensemble models

- Random Forest
- XGBoost with LambdaMART objective
- LGBM with LambdaMART objective

Features for Trees

- Features based on statistical language models
- ChatNoir relevance score (custom BM25 ranking function[4] based on inverted index)
- Comparative-based features

[3] `https://pyterrier.readthedocs.io/en/latest/index.html`

[4] `https://www.elastic.co/guide/en/elasticsearch/reference/current/`

Feature selection

Statistical features

- PyTerrier provides text features computed using the statistic language models (Tf, PL2...)
- To select three most informative, we rank document in validation set using every feature model

Table: 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

Feature selection

A comparative sentence has structures - objects for comparison, aspects and predicates. We apply the sequence-labelling model based on RoBERTa to the topic for defining comparative structures. Then we try to find them in the retrieved documents.

- `is_retrieved` describes are there any comparative structures in the document at all
- `objs_score` defines how many objects from topic are found in document
- `asp_pred_score` is counted in the following way: if at least one object from a topic is in the document, aspect or predicate increases the score to 0.5.

Re-ranker based on Bert

- We use reranking model from **OpenNIR**[5]. It is "Vanilla" Transformer architecture.
- We pre-train the model with ANTIQUE dataset. ANTIQUE contains 2,626 non-factoid questions from a diverse set of categories.
- We fine-tune the model with 40 topics from Train Set.

Table: Example of query and documents with different relevances 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

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

Results on Validation set

The best scores come from the LightGBM model, which also outperforms the baseline.

Table: 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	-

Feature importance

Feature importance in the proposed LightGBM model

Feature	Pl2	TF-IDF	BM25	Dfic	ChatNoir	is_retr	objs	asp_pred
Importance	1.76	1.19	1.51	2.3	20.8	0	1.66	1.51

Results on Test set

Table: 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: 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

The XGBoost model describes relevance a bit better and has first place in the table. LightGBM is better for quality and takes second place, slightly surrendering to Top 1.

Example output

Table: 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.

Conclusion

- We apply several approaches to the Argument retrieval shared task. We use ensembles-based methods and methods based on Transformer architecture.
- The best scores give gradient boosting models.
- Transformer-based model gives not very high performance. Perhaps this is due to the lack of relevant data for training.