MPII at the NTCIR-15 WWW-3 Task

Canjia Li
University of Chinese Academy of Sciences
licanjia17@mails.ucas.ac.cn

Andrew Yates
Max Planck Institute for Informatics
ayates@mpi-inf.mpg.de

ABSTRACT

MPII participated in the English subtask of WWW-3 at NTCIR-15 with several variants of our recent PARADE model. PARADE aggregates passage-level relevance representations into a document-level representation, which is then used to predict a document's relevance score. We submitted the best-performing PARADE variants in three runs. Our results support the findings in the PARADE paper: aggregating representations is more effective than aggregating scores, and effectiveness increases with the complexity of the aggregation approach.

TEAM NAME

MPII

SUBTASKS

English

1 INTRODUCTION

MPII participated in the English subtask of WWW-3 at NTCIR-15 in order to better evaluate variants of our PARADE ranking model [3]. As described in the official overview paper [6], this is an ad hoc retrieval task with queries and documents from the Web domain.

2 METHOD

We adopt the PARADE document reranking model [3], which has been shown to work well on standard TREC collections like Robust04 [7]. PARADE utilizes a pre-trained language model, such as BERT [2] or ELECTRA [1], as a building block for representing passages within a full document and learns independent passage relevance representation for a query-passage pair. The passage representations are aggregated using one of several approaches: PARADE $_{\text{Avg}}$, PARADE $_{\text{Max}}$, PARADE $_{\text{Attn}}$, or PARADE. More details about PARADE can be found in the original work [3]. We submitted runs corresponding to the aggregation approaches that previous performed best in the original work: Max, Attn, and the full Transformer-based model. The purpose of our WWW-3 submission was to compare these three aggregation strategies in a new and unbiased setting.

3 EVALUATION

3.1 Data

We split the documents into 32 passages using a sliding window of size 150 words with an overlap of 50 words. That is, 3250 words are preserved in each document for end-to-end document level optimization. The maximum sequence length in the pre-trained model is set to 256.

Run Name	PARADE variant	nDCG@10	Q@10	nERR@10
mpii-E-CO-NEW-3	PARADE $_{\text{Max}}$	0.6337	0.6556	0.7395
mpii-E-CO-NEW-2	PARADE $_{\text{Attn}}$	0.6743	0.6905	0.7787
mpii-E-CO-NEW-1	PARADE	0.6897	0.7016	0.8090

Table 1: Reranking effectiveness of PARADE.

3.2 Training

We used the ELECTRA-Base model to initialize PARADE. To prepare the model for the ranking task, we first fine-tuned the ELECTRA model on the MSMARCO passage ranking dataset.[1] We trained PARADE on the NTCIR WWW-1 and WWW-2 queries [4, 5], which are 180 in total. Given the top 100 results from the organizers' runs, documents labeled as relevant in the qrels are taken as positive examples; others including the unlabeled documents are taken as negative examples. Afterwards, the model is trained using a cross entropy learning objective. Training was conducted for 3 epochs with batches of 32 instances. We set the learning rate as 3e-6, with warm-up over the first 10 proportions of training steps. For both training and prediction, we reranked the top 100 documents provided by the official baseline. All experiments were conducted on a Google TPU v3-8.

3.3 Result

We submitted three runs for the English sub-task. The relationship between the run names and PARADE variants, as well as their effectiveness, are shown in Table 1. The trends observed mirror those observed by Li et al. [3], with effectiveness increasing as the complexity of the aggregation approach increases. However, as reported in the task overview [6], the differences between runs are not significant using Tukey's HSD test at the 5% significance level. As in the original work, the full PARADE model is the most effective across metrics. The PARADE $_{\text{Attn}}$ variant, which replaces PARADE's transformer encoder stack with a simple single-layer attention mechanism, performs slightly worse. The PARADE $_{\text{Max}}$ variant, which does not contain any new weights, performs the worst by a substantial margin. These results support the findings by Li et al. that aggregating passage representations is more effective than aggregating passage scores and that the full PARADE model is more effective than the simpler variants.

4 ANALYSIS

In this section, we conduct per-topic analysis to better understand the PARADE model.

[1]The fine-tuned ELECTRA model is available online at https://zenodo.org/record/3974431 and as part of the Capreolus implementation as `electra-base-msmarco`. [8] The original PARADE implementation available at https://github.com/canjiali/PARADE was used in our experiments.

4.1 System Comparison

To compare PARADE with the overall participant runs, we plot the per-topic gain or loss compared with the median score for all runs. As illustrated in Figures 1-3, PARADE outperforms the median for a vast majority of topics, followed by PARADE$_{\text{Attn}}$ and PARADE$_{\text{Max}}$. The observation is consistent for all metrics, which confirms the effectiveness of the full PARADE model.

4.2 Case Study

We further report the topics on which PARADE performs best or worst. We conduct cross-system comparison by setting t to the median nERR@10 and in-system comparison by setting $t = 0$. Then we rank the topics either descendingly to obtain the best-performing topics or ascendingly to obtain the worst-performing topics. As it can be seen from Tables 2-4, all PARADE variants fail for query *You want to visit the website "www.freeweblayouts.net"*. PARADE and PARADE$_{\text{Attn}}$ also fail for the topic that asks for a website: *You want to find the official website of Akron Beacon Journal* while PARADE$_{\text{Max}}$ suffers more from the query *You want to know how Zeus is described in the Greek Mythology.* For the queries seeking websites, regarded as known-item search, it might not be necessary to employ a full-document ranking model. The contextualization ability from pre-trained models introduces some noises for the ranking model while models adopt exact match features can be more reliable.

5 CONCLUSION

In the web ad-hoc ranking task, we confirm the effectiveness of PARADE. Results further support our finding that a better passage relevance representation aggregation approach makes up a more effective full-document ranking model.

REFERENCES

[1] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net. https://openreview.net/forum?id=r1xMH1BtvB

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1).* Association for Computational Linguistics, 4171–4186.

[3] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PARADE: Passage Representation Aggregation for Document Reranking. *CoRR* abs/2008.09093 (2020). arXiv:2008.09093 https://arxiv.org/abs/2008.09093

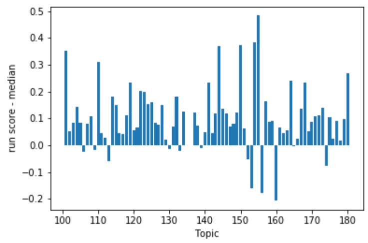
[4] Cheng Luo, Tetsuya Sakai, Yiqun Liu, Zhicheng Dou, Chenyan Xiong, and Jingfang Xu. 2017. Overview of the NTCIR-13 We Want Web Task. In *NTCIR-13.*

[5] Jiaxin Mao, Tetsuya Sakai, Cheng Luo, Peng Xiao, Yiqun Liu, and Zhicheng Dou. 2019. Overview of the NTCIR-14 We Want Web Task. In *NTCIR-14.*

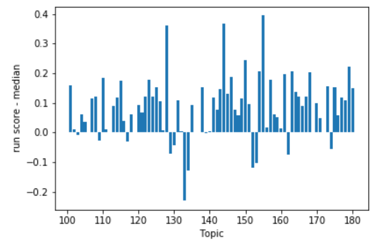
[6] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, and Ian Soboroff. 2020. Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task. In *Proceedings of NTCIR-15.* to appear.

[7] Ellen M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *TREC (NIST Special Publication)*, Vol. 500-261. National Institute of Standards and Technology (NIST).

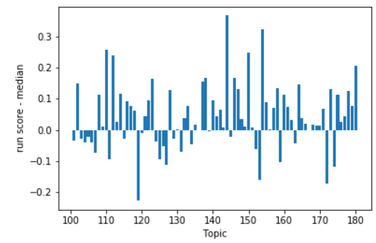
[8] Andrew Yates, Kevin Martin Jose, Xinyu Zhang, and Jimmy Lin. 2020. Flexible IR Pipelines with Capreolus. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.* 3181–3188.



(a) PARADE

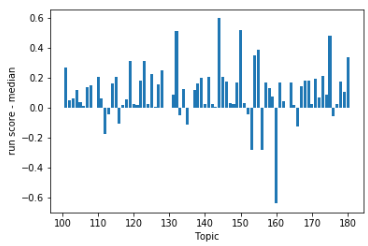


(b) PARADE$_{\text{Attn}}$

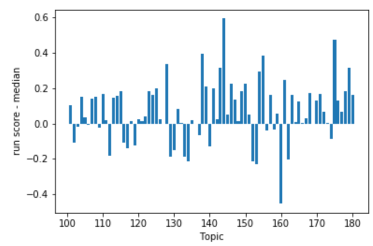


(c) PARADE$_{\text{Max}}$

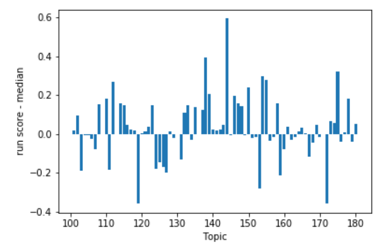
Figure 1: Per-topic difference from median nDCG@10 for all runs



(a) PARADE

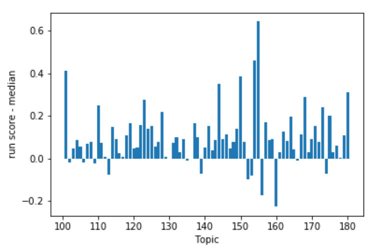


(b) PARADE$_{\text{Attn}}$

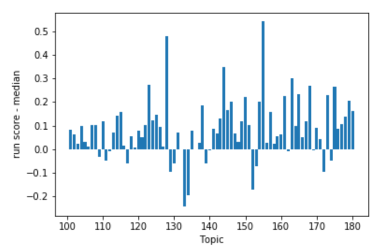


(c) PARADE$_{\text{Max}}$

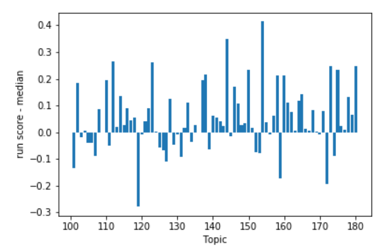
Figure 2: Per-topic difference from median nERR@10 for all runs



(a) PARADE



(b) PARADE$_{\text{Attn}}$



(c) PARADE$_{\text{Max}}$

Figure 3: Per-topic difference from median Q@10 for all runs

Type	QID	Score - t	Query
t = median	144	0.5942	Teacher's Day is approaching and you want to know its origin.
	150	0.5164	Your PC clock is currently inaccurate so you want to find out how to set it accurately.
	132	0.5085	Friends is a very famous TV series, you want to know in which year it first aired.
	160	-0.6407	You want to find the official website of Akron Beacon Journal.
	156	-0.2869	You are planning to buy a car and want to know the specs of Toyota Corolla.
	153	-0.2837	You want to visit the website "www.freeweblayouts.net".
$t = 0$	110	1.0000	You want to find out how to send a parcel by FedEx.
	121	1.0000	You want to learn how to treat spinal stenosis.
	157	0.9998	You are about to get married. As a groom, you want to know what traditional wedding vows are like.
	153	0.0000	You want to visit the website "www.freeweblayouts.net".
	147	0.2350	You want to know whether there are health benefits of bird nest.
	169	0.2702	You want to know when and where paper was invented.

Table 2: Queries solved best/worst by PARADE according to nERR@10.

Type	QID	Score - t	Query
t = median	144	0.5942	Teacher's Day is approaching and you want to know its origin.
	175	0.4775	You want to know the reputation of fort smith public schools.
	138	0.3958	You want to know the definition of Smart Home.
	160	-0.4523	You want to find the official website of Akron Beacon Journal.
	153	-0.2297	You want to visit the website "www.freeweblayouts.net".
	134	-0.2160	You are suffering from a shoulder pain and want to find out the symptoms of frozen shoulder.
$t = 0$	108	1.0000	You want to know how epilepsy can be treated.
	120	1.0000	You want to know what harm asbestos does to the human body.
	121	1.0000	You want to learn how to treat spinal stenosis.
	153	0.0540	You want to visit the website "www.freeweblayouts.net".
	132	0.0772	Friends is a very famous TV series, you want to know in which year it first aired.
	169	0.0901	You want to know when and where paper was invented.

Table 3: Queries solved best/worst by PARADE$_{\mathbf{Attn}}$ according to nERR@10.

Type	QID	Score - t	Query
t = median	144	0.5942	Teacher's Day is approaching and you want to know its origin.
	138	0.3928	You want to know the definition of Smart Home.
	175	0.3220	You want to know the reputation of fort smith public schools.
	119	-0.3594	You want to know how Zeus is described in the Greek Mythology.
	172	-0.3567	You want to know about the benefits of mineral essence to skin.
	153	-0.2837	You want to visit the website "www.freeweblayouts.net".
$t = 0$	121	1.0000	You want to learn how to treat spinal stenosis.
	116	0.9999	You want to start investing in stocks and need some advice from experts.
	117	0.9999	You are planning to buy a pageant dress at a shop, but now doing a research online to see what options there are.
	153	0.0000	You want to visit the website "www.freeweblayouts.net".
	169	0.1351	You want to know when and where paper was invented.
	132	0.1801	Friends is a very famous TV series, you want to know in which year it first aired.

Table 4: Queries solved best/worst by PARADE$_{\mathbf{Max}}$ according to nERR@10.