Entity-Conditioned Question Generation for Robust Attention Distribution in Neural Information Retrieval

Revanth Gangi Reddy[1], Md Arafat Sultan[2], Martin Franz[2], Avirup Sil[2], Heng Ji[1]

[1]University of Illinois at Urbana-Champaign [2]IBM Research AI

Motivation

- We observe that neural IR models can give low attention to many potentially important words and phrases in the passage, e.g, *academy of management* and *twentieth century*.
- This leads to relatively low IR scores for questions that are about these less-attended entities.
- In 65% of the cases, the highest-attended entity is present in the first half of the passage. The lowest-attended entity is in the second half in 60% of the cases.
- Such biases in IR models can be overcome by generating synthetic data that is targeted towards these shortcomings.

[CLS] frederick winslow taylor [SEP] frederick winslow taylor (march 20 , 1856 march 21 , 1915) was an american mechanical engineer who sought to improve industrial efficiency . he was one of the first management consultants . taylor was one of the intellectual leaders of the efficiency movement and his ideas , broadly conceived , were highly influential in the progressive era (1890s - 1920s) . taylor sum ##med up his efficiency techniques in his 1911 book " the principles of scientific management " which , in 2001 , fellows of the academy of management voted the most influential management book of the twentieth century . his pioneering work in applying engineering principles to the work [SEP]

Figure 1: Heatmap of attention given to each token in DPR's passage representation. Darker shading indicates more attention.

Question	Score
the *american mechanical engineer* who sought to improve *industrial efficiency*	85.9
who wrote the *most influential management book* of the *twentieth century*	78.0
who was considered the father of management during the *progressive* era	82.2
who wrote the *principles of scientific management*	86.8

Table 1: Retrieval scores from DPR for different questions corresponding to the passage in left. Important terms in the question, that are also in the passage, are shown in *italics*.

Contributions

- We introduce an entity-conditioned data augmentation strategy for IR, that generates questions about less-attended entities in the passage.
- We propose to incorporate these conditionally generated questions into the synthetic pre-training, to help improve model attention patterns and thereby the retrieval performance.

Entity-Conditioned Question Generation

- Given a passage and an entity in that passage, we aim to generate a synthetic question about that entity.
- While training the synthetic question generator, entities within questions in existing machine reading comprehension datasets are matched against the passage to identify the conditioning entities.
- While generating synthetic IR data, entities that get lowest attentions from the IR model are used as the conditioning entities.

Frederick Winslow Taylor PERSON (March 20 , 1856 DATE – March 21 , 1915 DATE) was an American NORP mechanical engineer who sought to improve industrial efficiency . He was one CARDINAL of the first ORDINAL management consultants . Taylor PERSON was one CARDINAL of the intellectual leaders of the Efficiency Movement ORG and his ideas , broadly conceived , were highly influential in the Progressive Era (1890s - 1920s) DATE . Taylor PERSON summed up his efficiency techniques in his 1911 DATE book " The Principles of Scientific Management " WORK_OF_ART which , in 2001 DATE , Fellows of the Academy of Management ORG voted the most influential management book of the twentieth century DATE .

Figure 2: Entities extracted from the passage shown in Figure 1.

Conditioned Entity	Generated Question
Progressive Era	who was considered the father of management during the progressive era
Principles of Scientific Management	who wrote the principles of scientific management
Efficiency Movement	who is known as the father of efficiency movement

Table 2: Questions output by the entity-conditioned generation system for the passage in Figure 2.

Overall Framework

- We first identify entities that received lowest attention in the passages, by aggregating word-piece level attentions from the last transformer layer into phrase-level attentions.
- Synthetic questions about the lowest-attended entities are generated from the passage, using the entity-conditioned question generator.
- We use a cycle-consistency filter with a question answering model to filter out low quality questions for which the passage doesn't answer the question.
- From the remaining synthetic questions, harder questions are retained based on question-passage IR scores from the baseline neural IR model.

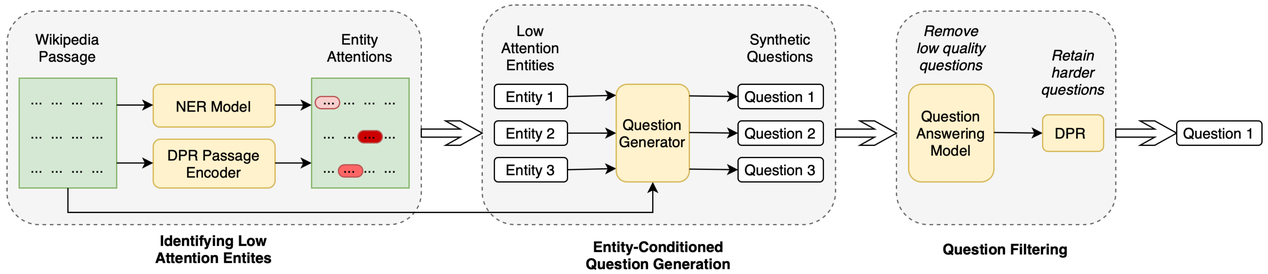

Figure 3: Overall framework of our synthetic data generation process to generate questions about named entities that receive low attentions from the DPR model.

Experiments

- The model that uses the entity-conditioned questions within its pre-training is named *Mixed-DPR*, and is compared with the baseline DPR.
- We also compare with a model pre-trained on data that contains synthetic questions generated without any conditioning (*UnCon-DPR*).
- We see that Mixed-DPR gives upto 2% more attention to latter sentences of the passage, compared to the baseline DPR model.
- Mixed-DPR also has the highest entropy (4.10) for attention over the passage tokens, compared to the baselines (3.97 for DPR, 3.80 for UnCon-DPR), meaning attention is more scattered.

Model	Natural Questions (NQ)						WebQuestions	
	Full test		No ans. overlap		No ques. overlap		Test	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
TF-IDF	14.2	32.0	13.6	28.6	14.6	31.8	14.5	32.1
BM25	22.7	44.6	20.1	39.6	24.0	43.4	18.9	41.8
DPR (ours)	44.3	67.1	32.2	53.2	37.2	60.1	29.4	51.6
UnCon-DPR	45.8	68.4	32.7	54.4	36.9	60.6	31.5	53.2
Mixed-DPR	**45.9**	**69.0**	**33.8**	**55.7**	**37.9**	**62.0**	**32.2**	**53.9**

Table 3: Top-***k*** retrieval results (in %) on test sets of Natural Questions and WebQuestions. Numbers on WebQuestions are in zero-shot settings, since models have been trained on NQ.