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

Anonymous ACL submission

Abstract

We show that supervised neural information retrieval (IR) models are prone to learning sparse attention patterns over passage tokens, which can result in key phrases including named entities receiving low attention weights, eventually leading to model under-performance. Using a novel targeted synthetic data generation method that identifies poorly attended entities and conditions the generation episodes on those, we teach neural IR to attend more uniformly and robustly to all entities in a given passage. On three public IR benchmarks, we empirically show that the proposed method[1] helps improve both the model's attention patterns and retrieval performance, including in zero-shot settings.

1 Introduction

Neural information retrieval (IR) performs query-passage matching at a semantic level, often using a dual-encoder architecture that encodes the queries and the passages separately. Examples of such models include the Dense Passage Retriever (DPR) (Karpukhin et al., 2020) and ANCE (Xiong et al., 2020), which fine-tune transformer-based (Vaswani et al., 2017) pre-trained language models (Devlin et al., 2019) to compute contextualized representations of queries and passages.

In this paper, we first uncover a shortcoming in the passage encoder of such a dual-encoder IR model, namely DPR, which stems from its sparse attention pattern. To illustrate, in Figure 1 we show a heatmap of the attention weights of DPR's passage encoder over different tokens of an example passage (taken from the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019)). We can see that the attention given to many potentially important words and phrases, e.g, *academy of management* and *twentieth century*, are rather low.

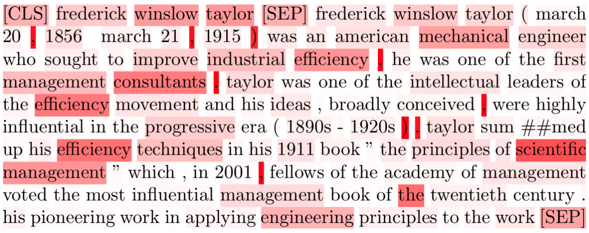

[CLS] frederick winslow taylor [SEP] frederick winslow taylor (march 20 , 1856 march 21 , 1915) was an american mechanical engineer who sought to improve industrial efficiency . he was one of the first management consultants . taylor was one of the intellectual leaders of the efficiency movement and 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.

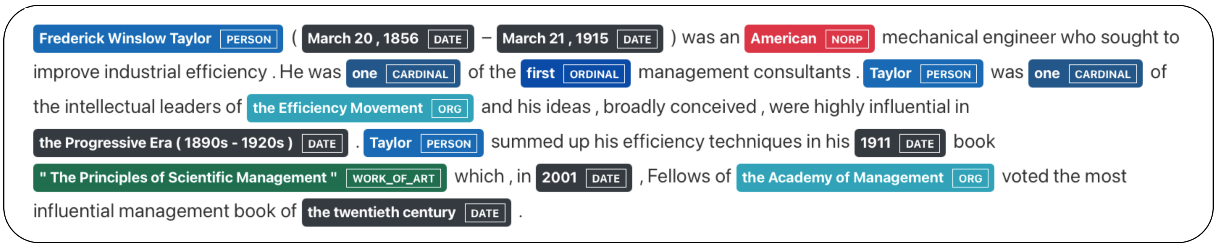
Question	Type	Score
the *american mechanical engineer* who sought to improve *industrial efficiency*	G	85.9
who wrote the *most influential management book* of the *twentieth century*	S	78.0
who was considered the father of management during the *progressive* era	S	82.2
who wrote the *principles of scientific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

Table 1: Retrieval scores from DPR for the passage in Figure 1, against both a gold-standard question (G) from NQ and three synthetic questions (S). The important terms in the question, that are also in the passage, are shown in *italic*.

What is the effect of such attention, or lack thereof, on retrieval performance? Table 1 shows DPR's retrieval scores for a gold-standard question (from the NQ dataset) and three automatically generated synthetic questions (details in Section 2) when paired with the passage of Figure 1. The gold-standard question, which overlaps highly with the well-attended first sentence of the passage, receives a relatively high retrieval score. Among the synthetic questions, the one that refers to the highest-attended entity (*principles of scientific management*) gets the highest score, whereas the ones about less attended entities (*twentieth century, progressive era*) receive considerably lower scores.

As models trained on limited amounts of human-labeled data are prone to biases such as these, here we also propose to augment the training data for

[1] We will make our code, data and models publicly available in the final version.



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 automatically extracted from the passage of Figure 1.

neural IR with synthetic questions that are conditioned on the sparsely-attended parts of the passage. Concretely, we generate questions specifically about entities that receive low attentions from the passage encoder of the neural IR model. Our experiments show that augmenting the training with such questions does indeed enable neural IR models to attend more uniformly over passage tokens, resulting in performance improvements on multiple benchmark datasets.

In contrast to existing work that unconditionally generate synthetic questions for tasks like question answering (Alberti et al., 2019; Sultan et al., 2020; Shakeri et al., 2020) and neural retrieval (Ma et al., 2021; Gangi et al., 2021; Reddy et al., 2021), our approach generates questions that are targeted towards the deficiencies of the neural IR model, by conditioning on the sparsely-attended entities in the passages.

Our main contributions are as follows:

- We show that a SOTA neural IR model is prone to learning sparse attention patterns over input passage tokens where key phrases (such as named entities) can receive low attention, leading to poor retrieval performance.
- We propose an entity-conditioned data augmentation strategy that generates questions about less attended entities in the passage.
- We demonstrate that incorporating these conditionally generated questions into the synthetic pre-training helps improve both model attention patterns and retrieval performance, including in zero-shot settings.

2 Method

To help neural retrievers capture all entities in the passage, we propose to augment the training data with synthetic questions that are conditioned on the less attended entities in the passage. Our synthetic data generation process involves the following steps: (a) Identifying entities with low attention, (b) Generating questions that are conditioned on these entities, and (c) Filtering out low-quality synthetic questions. We describe each step in detail

Identifying entities with low attention. We use a named entity recognition system to first identify all the entities in a given passage (see Figure 2). Then we compute attentions of the neural IR model over the passage and aggregate the attentions over the corresponding word-pieces to get the attention for each of the entities in the passage. Finally, we identify the entities with the lowest attentions.

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

Table 2: Questions output by the synthetic generation system for the passage in Figure 2, based on the entity used for conditioning.

Entity-conditioned question generation. Given a passage and an entity in that passage, we aim to generate a synthetic question about that entity using the passage. Specifically, we train a synthetic example generator to take a passage p, an entity e and generate a question q and its corresponding answer a. To achieve this, we fine-tune an encoder-decoder language model (Lewis et al., 2020a) using examples from existing machine reading comprehension (MRC) datasets, which take the form of (q, p, a) triples. Given such a triple, we first identify entities in q that also appear in p. One such entity e is passed as input along with p to condition the question generation. Following Sultan et al. (2020), we use top-p top-k sampling (Holtzman et al., 2020) during generation to promote sample diversity. Table 2 shows some generated questions conditioned on entities in the passage of Figure 2.

Question filtering. We employ a two-stage filtering process to promote high quality in the synthetic data. In the first stage, a generated question q is considered to be consistent with the input passage p if a separately trained MRC model can find an answer to q in p with high confidence. All other questions are filtered out. Among the remaining questions and their corresponding passages, we expect those to provide the best complementary signal (relative to existing gold-standard data) for which the baseline neural IR model has a low retrieval score. Hence, we only include such low scoring (harder) pairs in the synthetic pre-training set.

3 Experiments

3.1 Datasets

We use three public IR datasets in our experiments.

Natural Questions: We train all systems on Natural Questions (NQ) (Kwiatkowski et al., 2019), a dataset with questions derived from Google's search log and their human-annotated answers coming from Wikipedia articles. Lewis et al. (2020b) report that 30% of the NQ test set questions have near-duplicate paraphrases in the training set and 60–70% of the test answers are also present in the training set. For this reason, in addition to the original 3,610 test questions, we also report evaluation on the non-overlapping subsets (1,313 no-answer overlap and 672 no-question overlap) released by Lewis et al. (2020b).

TriviaQA: This dataset contains questions created by trivia enthusiasts from trivia and quiz league websites (Joshi et al., 2017). We use its 11,313 test questions for zero-shot evaluation.

WebQuestions: The dataset consists of questions obtained using the Google Suggest API, with answers selected from entities in Freebase by AMT workers (Berant et al., 2013). We use the 2,032 test questions in this dataset for zero-shot evaluation.

3.2 Setup

We use the 21M Wikipedia passages from Karpukhin et al. (2020) as the retrieval corpus for all our experiments.

Synthetic Data Generation. To create our synthetic pre-training corpus, first we derive a random sample of passages from the above collection. We identify the named entities in these passages using a publicly available NER system[2] trained on the OntoNotes corpus (Weischedel et al., 2011). We then fine-tune BART (Lewis et al., 2020a) for *conditioned* generation, which takes a (passage, entity) pair as input and generates an entity-conditioned question and its answer as output. This model is trained with examples from the NQ dataset. To obtain the conditioning entities used in training, we identify entities from noun chunks (obtained using spaCy (Honnibal et al., 2020)) in the question that also occur in the corresponding passage.

To compare our approach with a generation strategy that does not use any conditioning, we also train an unconditioned generation system, similar to Reddy et al. (2021), that generates (question, answer) pairs using just the passage as input. We use this generator to generate 1M synthetic examples, which we call *unconditioned* synthetic data.

We use the conditioned generation system to obtain 500k examples after filtering, and mix them with 500k unconditioned examples to obtain our final dataset of size 1M, which we call *mixed* synthetic data. Since the conditioned data contains questions primarily about less attended entities, this combination with unconditioned examples helps maintain adequate diversity in the final mixed dataset. We follow the same process as in Karpukhin et al. (2020) and use term matching to sample hard negatives for the questions.

Baselines. As a traditional term matching baseline, we evaluate the TF-IDF system[3] from Chen et al. (2017). We also evaluate DPR[4] as our neural IR baseline[5]. Karpukhin et al. (2020) report that the performance of DPR is affected by the number of in-batch negatives used in training, which in turn is dependent on the number of GPUs available. They use 128 in-batch negatives with eight 32GB V100s. Since we only had access to four 32GB V100s, we use 64 in-batch negatives. We call this model *DPR (ours)*, which we train on NQ for 40 epochs following Karpukhin et al. (2020).

Training. We pre-train both of our synthetically augmented DPR models for 10 epochs. We name

[2] https://demo.deeppavlov.ai

[3] https://github.com/efficientqa/retrieval-based-baselines#tfidf-retrieval

[4] https://github.com/facebookresearch/DPR

[5] We note that our approach can be similarly applied to other dual-encoder IR models like ANCE (Xiong et al., 2020).

Model	Natural Questions (NQ)						TriviaQA		WebQuestions	
	Full test		No ans. ovlp.		No ques. ovlp.		Test		Test	
	Top-1	Top-5	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	31.7	51.2	14.5	32.1
DPR (ours)	44.3	67.1	32.2	53.2	37.2	60.1	37.2	55.7	29.4	51.6
UnCon-DPR	45.8	68.4	32.7	54.4	36.9	60.6	37.6	57.2	31.5	53.2
Mixed-DPR	**45.9**	**69.0**	**33.8**	**55.7**	**37.9**	**62.0**	**38.3**	**57.5**	**32.2**	**53.9**

Table 3: Top-k retrieval results (in %) on test sets of Natural Questions (including the non-overlapping subsets of Lewis et al. (2020b)), TriviaQA and WebQuestions. Numbers on TriviaQA and WebQuestions are in zero-shot settings, since models have been trained on NQ.

the model pre-trained on the unconditioned synthetic data as *UnCon-DPR* and the one pre-trained on the mixed synthetic data as *Mixed-DPR*. After pre-training, both models are fine-tuned on NQ for 40 epochs. We refer the reader to the appendix for more details on hyper-parameters.

3.3 Results

Similar to Karpukhin et al. (2020), we evaluate all systems using top-k retrieval accuracy, which is the percentage of questions with at least one answer in the top k retrieved passages. Table 3 shows the results for the term matching and neural models.

Firstly, we can see that the two DPR models with synthetic pre-training improve over the baseline DPR system. Our Mixed-DPR model, which employs entity-conditioned synthetic questions for pre-training, consistently outperforms all other models including UnCon-DPR, which is pre-trained only on unconditioned questions. Crucially, on NQ, we observe greater improvements with Mixed-DPR on the non-overlapping and thus harder subsets of NQ, which indicates that the robustness of DPR improves with our proposed data augmentation strategy.

Analysis. To investigate the effect of the entity-conditioned questions used in synthetic pre-training, we examine how their application changes the attention distribution of DPR. First we observe that the gold-only DPR model tends to attend more to the earlier sentences of a given passage. We therefore compare attention on the first sentence (computed as the average attention over its tokens) with average attention on the rest of the sentences in the passage. We sample 10k passages from the retrieval corpus and compute attentions for the baseline DPR, UnCon-DPR and Mixed-DPR models. We observe that Mixed-DPR pays 1.8% higher attention to the later sentences of the passage compared to the baseline DPR model. When compared to UnCon-DPR, this difference is 1.1%. These results show that Mixed-DPR learns to attend more to the latter sentences of the passage which, as shown in Figure 1, is typically where most of the weakly attended entities of the baseline model occur.

Next, we look at the entropy of token-level attentions in a given passage for the above models. Entropy here is a measure of the uniformity of a model's attention over the tokens in the passage, with a higher entropy indicating a more uniform distribution. For the 10k passages previously sampled, we see that the baseline DPR, UnCon-DPR and Mixed-DPR models have attention entropies of 3.97, 3.80 and 4.10 respectively, with Mixed-DPR being the highest. This suggests that the improvements in top-k retrieval accuracy stem (at least partly) from a more scattered and potentially more robust attention pattern learned by Mixed-DPR.

4 Conclusion

We discover a specific issue in neural IR systems that stem from sparse attention patterns learned over input passage tokens, which can lead to suboptimal performance on queries about less attended areas of the passage. With targeted synthetic data augmentation, we address this issue for DPR—a state-of-the-art neural IR model—and enable it to attend more uniformly over passage tokens. Our proposed method improves the performance of DPR on three different benchmarks. While our work is an important first step towards solving this problem, one of our primary goals in this paper is to draw attention of the community to this important limitation of supervised neural IR and inspire future research on the topic. One potential direction is to incorporate additional objectives, e.g. multitask learning, to help models learn more robust attention patterns without requiring synthetic data.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Revanth Gangi Gangi, Bhavani Iyer, Md Arafat Sultan, Rong Zhang, Avirup Sil, Vittorio Castelli, Radu Florian, and Salim Roukos. 2021. Synthetic target domain supervision for open retrieval qa. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1793–1797.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020b. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088.

Revanth Gangi Reddy, Vikas Yadav, Md Arafat Sultan, Martin Franz, Vittorio Castelli, Heng Ji, and Avirup Sil. 2021. Towards robust neural retrieval models with synthetic pre-training. *arXiv preprint arXiv:2104.07800*.

Siamak Shakeri, Cicero dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for qa. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

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

Anonymous ACL submission

1 Appendix

1.1 Hyperparameters

In this section, we share the hyperparameters details for our experiments. Table 1 gives the hyperparameters for training the synthetic generator. These are the same for both the entity-conditioned and unconditioned synthetic generator.

Hyperparameter	Value
Learning rate	3e-5
Epochs	3
Batch size	24
Max Sequence length	1024

Table 1: Hyperparameter settings during training the synthetic question generator (BART) using data from Natural Questions.

Table 2 lists the hyperparameters for pre-training and finetuning the neural IR model.

Hyperparameter	Pre-training	Finetuning
Learning rate	1e-5	2e-5
Epochs	10	40
Batch size	1024	128
Gradient accumulation steps	8	1
Max Sequence length	256	256

Table 2: Hyperparameter settings for the neural IR model during pre-training on synthetic data and finetuning on NQ.

The MRC model used in the first-stage of question filtering is trained sequentially on SQuAD2.0 and Natural Questions, with hyperparameters shown in Table 3.

Hyperparameter	SQuAD2.0	Natural Questions
Learning rate	3e-5	2e-5
Epochs	3	1
Batch size	32	32
Max Sequence length	384	512

Table 3: Hyperparameter settings for training the MRC model on SQuAD2.0 and Natural Questions.