AutoQIR: Auto-Encoding Questions with Retrieval Augmented Decoding for Unsupervised Passage Retrieval and Zero-shot Question Generation

Stalin Varanasi[1,2] **Muhammad Umer Butt**[1] **Günter Neumann**[1,2]
[1]Saarland Informatics Campus, D3.2, Saarland University, Germany
[2] German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
stalin.varanasi@dfki.de, mubu00001@uni-saarland.de, guenter.neumann@dfki.de

Abstract

Dense passage retrieval models have become state-of-the-art for information retrieval on many Open-domain Question Answering (ODQA) datasets. However, most of these models rely on supervision obtained from the ODQA datasets, which hinders their performance in a low-resource setting. Recently, retrieval-augmented language models have been proposed to improve both zero-shot and supervised information retrieval. However, these models have pre-training tasks that are agnostic to the target task of passage retrieval. In this work, we propose Retrieval Augmented Auto-encoding of Questions for zero-shot dense information retrieval. Unlike other pre-training methods, our pre-training method is built for target information retrieval, thereby making the pre-training more efficient. Our method consists of a dense IR model for encoding questions and retrieving documents during training and a conditional language model that maximizes the question's likelihood by marginalizing over retrieved documents. As a by-product, we can use this conditional language model for zero-shot question generation from documents. We show that the IR model obtained through our method improves the current state-of-the-art of zero-shot dense information retrieval, and we improve the results even further by training on a synthetic corpus created by zero-shot question generation.

1 Introduction

Open Domain Question Answering (ODQA) with dense passage retrieval has been quite successful in recent years. This is primarily because of the availability of large question-answering corpora. However, annotations for the creation of Open-Domain Question Answering (ODQA) datasets consume significant time and effort, although, the indispensable need for labeled data is evident in the decline of cross-domain performance across various ODQA datasets (Karpukhin et al., 2020) for both information retrieval and question-answering tasks. To this end, in this work, we address the task of Unsupervised Dense Passage Retrieval (UDPR). That is, to be able to retrieve relevant documents without the labels on ground truth question-passage pairs, which reflects a real-world scenario.

In this work, we propose *Retrieval Augmented Auto-Encoding of Questions* (named as AutoQIR[1]) as a means to obtain similarity between documents and questions to perform zero-shot dense passage retrieval. Our method not only complements the supervised methods but unlike other zero-shot pre-trained models, it also considers a pre-training task that is directly relevant to *questions*. The following are the contributions of this work:

1. We propose a novel pre-training task for Unsupervised Dense Information Retrieval.
2. We provide a new method for zero-shot question generation which can be used for data augmentation of Question Answering/ IR Datasets.
3. We provide a way to transfer knowledge from language models to Information Retrieval.
4. Our method surpasses the baseline and is on par with other zero-shot dense information retrieval approaches. Additionally, our pre-training method is effective even with few thousand datapoints.

2 Related Work

Traditionally, lexical models with sparse vector spaces, such as BM25 (Robertson et al., 2009), have been used for unsupervised retrieval of the neighboring documents of a query. These models

[1] Auto-Encoding of Questions for Information Retrieval

consider documents and queries as bags of words and rely upon possible word overlap between the query and the relevant document to assign a high cosine similarity between them. Consequently, they suffer from the problem of the lexical gap between query and document (Berger et al., 2000) and are unable to capture the meaning that comes through word order.

Alternatively, Dense passage[2] retrieval models capture the meaning by encoding the sequence of words. Hence, unsupervised methods using dense passage retrieval can potentially yield better recall than the sparse retrieval models. Recently, several pre-training methods have been proposed to improve the joint dense embedding space of queries and documents. Retrieval augmented pre-training and fine-tuning methods (Guu et al., 2020; Lewis et al., 2020c,a) have been shown to improve dense passage retrieval. These methods train an information retrieval model to improve the context required for adjoining pre-training tasks. Amongst these, Guu et al. (2020) showed the ability for unsupervised dense passage retrieval while pre-training on Masked Language Modeling. Izacard et al. (2022) used a contrastive loss to discriminate between positive and negative documents while considering pseudo questions. While these methods are effective, the pre-training task chosen in their approach lacks explicit adaptation for the target task of passage retrieval for queries.

Estimating the likelihood of the *question* given a context is useful in various steps of Question Answering and Information Retrieval tasks. For example, (Lewis and Fan, 2018) maximized question likelihood (by decomposing the posterior probability) instead of answer likelihood and showed that QA models relying on question likelihood are robust to perturbations in the input. Another approach (Lewis et al., 2019) used unsupervised question generation methods to augment data for extractive question answering. Varanasi et al. (2021) used auto-encoding of questions for unsupervised answer span selection. It is shown by Sachan et al. (2022) that pre-trained language models can be used to re-rank the retrieved documents via 'prompt-based' question likelihood. Furthermore, parallel to our work, Sachan et al. (2023) have proposed that the retrieved documents (Lewis et al., 2020c) can be used to finetune a retriever by a teacher-student network. In their approach, the ground-truth distribution of the documents given a question is derived from the output of a frozen large pre-trained language model ($>$ 3B parameters). The dense retriever is trained by minimizing the KL divergence between its estimated distribution with the aforementioned ground truth distribution of the teacher network. The main difference between our work and theirs is that we utilize an auto-encoding loss while fine-tuning a BART decoder (406M parameters), thereby avoiding sole reliance on pre-existing (large) language models. Consequentially, our model can perform zero-shot question generation in addition.

3 Approach

Maximizing the likelihood of *question* given a context has been proven useful for Information Retrieval and Question Answering tasks (Zhao et al., 2021; Lewis and Fan, 2018; Nogueira et al., 2019). However, in an unsupervised setup, we don't have access to ground truth questions associated with passages. To mitigate this, we propose auto-encoding of questions by assuming an underlying conditional distribution over documents. In other words, our approach seeks to maximize the likelihood of a question by first obtaining the relevant passages. For this, we take the approach proposed by Lewis et al. (2020c).

Our training setup requires a set of questions Q and a set of documents S and no further labels for *answers* or *relevant documents*. Note that both sets of Q and S can be obtained without human annotations, for example, via web crawling. Our only assumption is that the set S contains relevant documents to most of the questions in set Q. Without this assumption, we model a uniform conditional distribution over documents. This expectation of relevant documents in a document corpus is fairly common in situations where an information retrieval task ought to be performed.

Formally, we aim to reconstruct the input question by assuming *document* z, as a latent variable. The loss L is obtained as the negative log-likelihood of the reconstructed question $\hat{q}$ given the input question q, as shown in eq. 1. The probability $p(\hat{q}|q)$ can be further decomposed by marginalizing over all known documents in the corpus S as shown in eq. 2. The input q for the conditional language model may provide an unwanted strong signal during reconstruction. This will lead to over-fitting of

[2] Please note that we use the terms *document* and *passage* interchangeably throughout this paper

the decoder and a weak encoder. Hence, we relax this term to $p(\hat{q}|z_i)$ by removing the dependency on input question q. Furthermore, the sum in eq. 2 is intractable to compute especially when the set S is very large. Also, note that when S is very large, most of the documents will have probabilities close to zero. To mitigate this, we approximate the sum by taking top-k documents.

Similar to Lewis et al. (2020c), our method mainly consists of two components: a *passage retriever* and a *sequence-to-sequence generator*. The equations below describe our loss function:

$$L = -\sum_{q \in Q} logp(\hat{q}|q) \qquad (1)$$

$$p(\hat{q}|q) = \sum_{z_i \in S} p(\hat{q}|q, z_i) * p(z_i|q) \qquad (2)$$

$$p(\hat{q}|q) \approx \sum_{z_i \in topk(q,S)} p_\phi(\hat{q}|z_i) * p_\theta(z_i|q) \qquad (3)$$

Eq. 3 above, describes our final model. $p_\theta(z_i|q)$ is a information retrieval model (*passage retriever*), $p_\phi(\hat{q}|z_i)$ is a conditional language model (*sequence-to-sequence generator*). θ and ϕ are the model parameters. In practice, the top-k documents are obtained during training by the information retrieval model $p_\theta(z_i|q)$.

3.1 Passage Retriever

Passage retriever is an information retrieval module that comprises of two encoders, one to encode *question* and the other to encode *document*. These encoders provide a dense embedding given an input text and by using dense embeddings, we keep this module differentiable. Similar to DPR (Karpukhin et al., 2020), we model these encoders as BERT[3] transformer models. Following standard practices, we provide BERT an input text prepended with '[CLS]' and post-pended with a '[SEP]' token. The output embedding of BERT at the position of [CLS] token is considered as the embedding of an input sequence x. We represent this by $BERT(x)$. We obtain the probability $p(z_i|q)$ as follows:

$$\vec{z_i} = W_{doc} BERT_{doc}(z_i)$$

$$\vec{q} = W_q BERT_q(q)$$

$$sim(z_i, q) = e^{<\vec{zi},\vec{q}>}$$

$$p(z_i|q) = \frac{sim(z_i, q)}{\sum_{z_j \in topk(q,S)} sim(z_j, q)} \qquad (4)$$

where W_q and W_{doc} are matrix parameters. Equation eq. 4 refers to the *sofmax* function applied on the similarity scores of question-document pairs. For retrieving top-K documents related to the question q, we use maximum inner-product search (MIPS) to obtain the 'k' nearest neighbors of the question embedding $\vec{q}$ in the set of documents S. During training, we use an indexed set of documents for fast retrieval[4].

3.2 Sequence-to-Sequence Generator

Given top-k relevant passages for a question q, the sequence-to-sequence generator estimates the likelihood of q given each passage using a transformer-based encoder-decoder mechanism (Vaswani et al., 2017) which we initialize using the pre-trained weights of BART-large model with 406M parameters. We estimate the probability of the question $\hat{q}$ as a product of probabilities of individual tokens similar to (Lewis et al., 2020c) as follows:

$$p(\hat{q}|q) = \Pi_{i=1..|\hat{q}|} \sum_{z_i \in topk(q,S)} p_\phi(\hat{q_j}|z_i, \hat{q}_1..\hat{q}_{j-1}) * p_\theta(z_i|q) \qquad (5)$$

4 Implementation Details

4.1 Initialization

The *passage retriever* and *sequence-to-sequence generator* are optimized during the training. However, a good initialization of *passage retriever* is required to obtain relevant passages during the initial stages of the training. We consider the following pre-trained models (which are also unsupervised) for initializing 'passage retriever'.

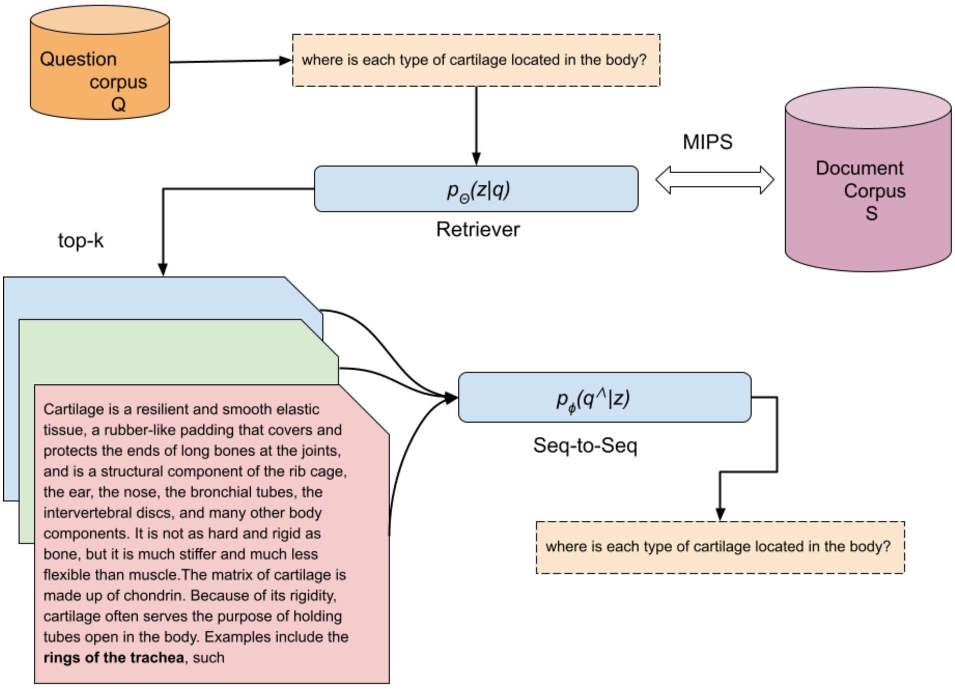
ICT: To obtain this initialization, we first train a dense passage retriever model (DPR) (Karpukhin et al., 2020) with the Inverse-Cloze Task (Lee et al., 2019) using a pseudo Question Answering Corpus of 100k data points. We further pretrain on the same dataset using the AutoQIR model with the missing sentences as pseudo questions.

REALM[5]: is the pretrained model proposed by Guu et al. (2020). This is one of the first dense retrieval models to show zero-shot abilities.

[3] We used uncased model with 110M parameters

[4] We use FAISS search on indexed document embeddings

[5] https://huggingface.co/docs/transformers/model_doc/realm


Question corpus Q
where is each type of cartilage located in the body?
MIPS
Document Corpus S
$p_\Theta(z|q)$
Retriever
top-k
$p_\phi(q^\wedge|z)$
Seq-to-Seq
where is each type of cartilage located in the body?

Cartilage is a resilient and smooth elastic tissue, a rubber-like padding that covers and protects the ends of long bones at the joints, and is a structural component of the rib cage, the ear, the nose, the bronchial tubes, the intervertebral discs, and many other body components. It is not as hard and rigid as bone, but it is much stiffer and much less flexible than muscle.The matrix of cartilage is made up of chondrin. Because of its rigidity, cartilage often serves the purpose of holding tubes open in the body. Examples include the **rings of the trachea**, such

Figure 1: Overview of *Retrieval augmented Question Auto-Encoding*. The *Retriever* module retrieves top-k documents for each input question using maximum inner-product search (MIPS) during training. Each of these documents is passed as input to the sequence-to-sequence module while reconstructing the input question.

4.2 Training

We initialize our sequence-to-sequence generator to BART (Lewis et al., 2020b) weights. We optimize for the loss mentioned in equation eq. 1. We take the value of k as 5 (in top-k documents) in our experiments. We optimize the question encoder of the passage retriever and *freeze* the weights for the context encoder to avoid refreshing indices at regular intervals as done by (Guu et al., 2020). We build the index of all candidate documents before beginning the training.

The training is terminated using early stopping when the training objective plateaus on the validation set. The training is performed on a Tesla V100 GPU with 32GB RAM and a batch size of 4. [6]

During inference, we discard the sequence-to-sequence model and use only the ’passage retriever’ module for retrieving documents.

4.3 Datasets

We use 5 commonly used datasets for open domain question answering: SQuAD (Rajpurkar et al., 2016), Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), Web Questions (Berant et al., 2013), Curated Trec (Baudiš and Šedivỳ, 2015). We trained separately on multiple-question corpora (Q) and corresponding multiple-document corpora (S). The question corpora (Q) is formed from the questions of the training sets of the aforementioned datasets. We use a segmented Wikipedia corpus provided by Karpukhin et al. (2020), comprising approximately 21 million documents. Each passage in this corpus consists of 100 words, effectively serving as our document corpus (S) for the task. As mentioned in section 3, the corpus S is expected to contain answers for questions in Q. This expectation is met since the contexts for the questions in the aforementioned datasets are sourced from Wikipedia. Nevertheless, during training, the retrieval of top-k documents from such a large corpus can be significantly time-consuming. To speed this up, we split the question corpus into multiple sets of 1000 questions each, and the top 1k passages for each question in the corpus (S) are taken using bm25 to form the corresponding document corpus (S) (i.e., limiting the size of the document corpus to 1 million documents per set). During inference, we use the same segmented Wikipedia corpus for passage retrieval.

[6] For NQ questions, it takes around 35 hours to train for 7 epochs.

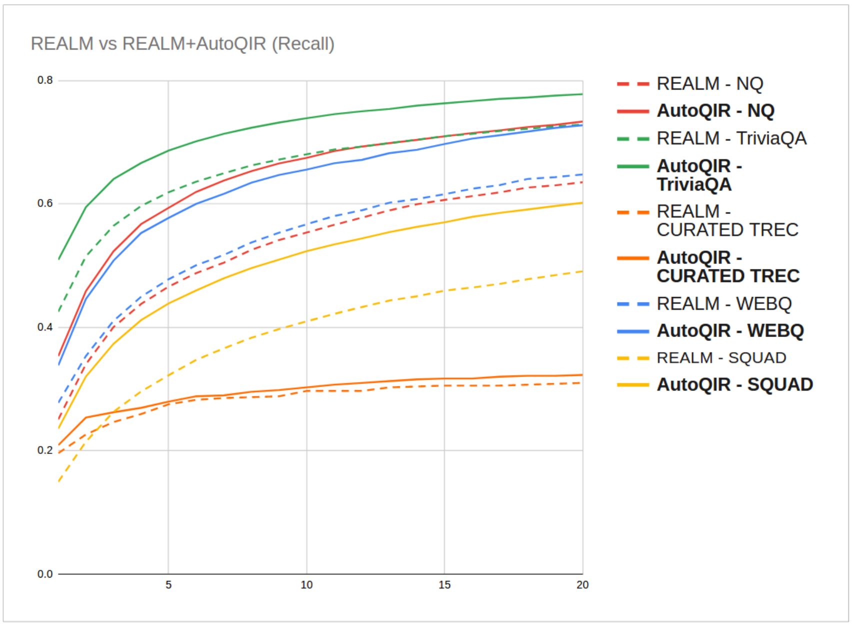


Figure 2: Comparison of the performances of REALM and $AutoQIR_{REALM}$ on recall@1 to recall@20 on various datasets. The dotted lines indicate REALM and stronger lines indicate $AutoQIR_{REALM}$ models.

5 Experiments

5.1 Main Results

In this section, we show that retrieval augmented auto-encoding of questions by itself is a useful tool for unsupervised information retrieval. We use *recall at top-k* (recall@k) as our evaluation metric as it reliably correlates with the information retrieval capabilities of the model.

Firstly, we observed the improvements of AutoQIR over the initialized baseline models - ICT and REALM. In table 1, it can be seen that AutoQIR models consistently outperform their corresponding initial models on various datasets by a big margin. Please note that AutoQIR models are trained on the set of questions from the corresponding training set mentioned in the columns. The AutoQIR models initialized with ICT pre-training, as mentioned in section 4, performs comparably to the baseline REALM model on all datasets except on the dataset CuratedTREC. This could be because of the low number of training samples available for this dataset. Whereas the AutoQIR models initialized with the REALM model improve the average recall@1 of REALM to 6.7 points across the 5 datasets. The importance of auto-encoding questions over auto-encoding sentences (ICT) can be seen from the contrasting differences in the results of $AutoQIR_{ICT}$ and ICT. In our experiments, we found that optimizing the decoder is more effective than using a frozen pre-trained language model as decoder. Figure 2 shows the comparison between the performance of initialized REALM model and its AutoQIR pre-training across all datasets for recalls between 1 and 20. AutoQIR consistently outperforms the baseline REALM model for all recalls with a large margin on all datasets except for CuratedTREC.

In table 2, we compare our best model with state-of-the-art unsupervised retrieval models. *Contriever* is a dense passage retriever model trained with a contrastive loss on a pseudo-question answering dataset. *Masked Salient Spans* model is also a dense passage retrieval model trained on ”cloze” questions (sentences with masked salient spans such as named entities) similar to pre-training data of REALM (Guu et al., 2020). Unlike AutoQIR, both of these models use supervised training methods, albeit, on a pseudo corpus that can be obtained without annotations. BM25 is a lexical-based sparse retrieval model. REALM is the only other retrieval-augmented model which can be compared for zero-shot information retrieval for question answering. AutoQIR models outperform all

	NQ	**TriviaQA**	**SQuAD**	**WebQ**	**CuratedTREC**
Baselines					
ICT	6.59	11.15	6.88	8.7	5.18
REALM	25.19	42.51	14.97	27.75	19.59
Our models					
$AutoQIR_{ICT}$	24.32	37.77	17.99	23.67	2.16
$AutoQIR_{REALM}$	**35.05**	**50.09**	**23.56**	**33.80**	**20.89**

Table 1: Improved baseline: Recall@1 on test-sets for various datasets.

	NQ			**TriviaQA**		
	@5	@20	@100	@5	@20	@100
BM25 (Ma et al., 2021)	–	62.9	78.3	-	76.4	**83.2**
Masked salient spans (Singh et al., 2021)	41.7	59.8	74.9	53.3	68.2	79.4
Contriever(Izacard et al., 2022)	47.8	67.8	**82.1**	59.4	74.2	**83.2**
REALM (Guu et al., 2020)	45.7	61.8	74.9	61.8	72.8	80.6
$AutoQIR_{REALM}$ **(ours)**	**57.7**	**71.8**	81	**67.6**	**77.1**	**83.2**
DPR(Karpukhin et al., 2020) (supervised)	-	78.4	85.4	-	79.4	85.0

Table 2: $AutoQIR_{REALM}$ vs state-of-the-art unsupervised retrieval models: Recall@(5,20,100) on NQ and TriviaQA tests. Results on a supervised method (DPR) is provided for reference.

Models	#questions	**NQ**	**TriviaQA**	**SQuAD**	**WebQ**	**CuratedTREC**
REALM	-	54.46	68.03	40.96	56.69	29.68
$AutoQIR_{REALM}$						
NQ	(58k)	**67.45**	69.94	48.24	65.40	29.68
TriviaQA	(60k)	61.49	**73.87**	49.33	65.60	**30.83**
SQuAD	(78k)	62.63	70.93	**52.33**	**66.78**	30.11
WebQ	(3k)	59.66	70.31	45.67	65.55	30.40
CuratedTREC	(1k)	58.50	71.08	46.32	65.20	30.25

Table 3: $AutoQIR_{REALM}$ trained with questions from various datasets (rows) and corresponding retrieval results (recall@10) across all datasets (columns).

	NQ	TriviaQA	SQuAD	WebQ	CuratedTrec
REALM	30.22	32.44	12.82	19.49	**11.24**
$AutoQIR_{REALM}$	**35.57**	**32.91**	**13.94**	**20.52**	10.52

Table 4: We compare the Exact match score of a trained Question-Answering module for different retrievers with top-100 retrieved documents.

the aforementioned models, including bm25, for recalls 10 and 20 on NQ. For recall at top-100 documents, in the TriviaQA dataset, it can be seen that all models perform decently and close to each other. In the NQ dataset, Contriever performs only slightly better than our best model for recall@100. These results suggest that our model is a viable alternative to the state-of-the-art methods.

5.1.1 Cross-domain Questions

Considering the significance of questions over other types of sentences of auto-encoding, it would be interesting to see how AutoQIR performs across various domains. i.e., a model trained on one domain and evaluated on the other. The questions from these datasets vary in their distribution due to the differences in purposes and methods of collecting these datasets. In table 3, we show the cross-dataset retrieval performance of AutoQIR models. The large datasets (where we used more than 50 questions for training), i.e., TriviaQA, SQuAD, and NQ have the best performances when they are trained on the same domain. For

Model	Recall@1
$AutoQIR_{REALM}$	35.05
$AutoQIR_{REALM}$+ data-augmented fine-tuning	**37.08**

Table 5: Improved recall@1 on NQ dataset with additional training on a synthetic corpus as specified in section 5.2

smaller datasets, CuratedTREC and WebQ, models trained on SQuaD and TriviaQA respectively had the highest performance. This could be due to their lower number of training samples. It can be observed from the table that any form of AutoQIR training improved the results from the baseline REALM model. For example, the AutoQIR model trained with around 1k questions from the CuratedTREC dataset outperforms the REALM model on all datasets.

5.1.2 Question-Answering

Finally, to see how the retrieved documents are used for the subsequent task of Question Answering, we use a fully supervised "reader" model[7] provided by Karpukhin et al. (2020) and apply on the top-100 retrieved documents. The results can be seen in table 4. Our model brings 5 points of improvement on the Exact Match for the NQ dataset and marginal improvements on the rest of the datasets. This could be because of the increased recall at larger values of k for all the models (as also observed in table 2).

5.2 Zero-shot Question Generation

Once the AutoQIR model is trained, the sequence-to-sequence generator can be used for zero-shot passage-to-question generation (without a specific answer phrase). This is due to the fact that the sequence-to-sequence generator models $p(q|z_i)$ where z_i is a passage from the document corpus S.

Paragraph-level question generation can not be evaluated directly by measuring the similarity to ground truth questions (for example, via BLEU score) due to the variance in the distribution of questions that can be asked from the paragraph. Here, we evaluate the generated questions by measuring their use to information retrieval.

We use our best model $AutoQIR_{REALM}$ trained on the NQ dataset for zero-shot question generation. We use 50 thousand randomly chosen paragraphs from Wikipedia segmented to a length of 100 tokens as our input corpus. We generated one question for each of these input passages using beam search. We further take negative paragraphs by choosing one among the top-3 passages closer to the question using **bm25** (excluding the input passage). Since passages usually contain unique information, we expect that the top-3 retrieved passages often do not contain the answer even though quite close to the question. Hence these provide a better challenge for the Passage Retrieval model than using random passages which can be quite distant from the generated questions. We trained a fully supervised model (Karpukhin et al., 2020) on this dataset. This model further outperforms our best AutoQIR model by 2 points for recall at top-1 (recall@1) shown in table 5. Zero-shot Question Generation has larger applications in the field of Question Answering which can be explored in future work.

6 Conclusion

In this work, we propose a novel pre-training task to perform unsupervised information retrieval. Our method, which is based on *Retrieval Augmented Generation* (Lewis et al., 2020c), shows significant improvements from the baseline zero-shot retrieval models (ICT and REALM). Our cross-domain evaluation reveals the significance of using target questions for pre-training. We also show that auto-encoding on questions has a much greater impact than auto-encoding of sentences (ICT). Our model explicitly captures knowledge stored in language models into IR models. Additionally, our method can be used for zero-shot question generation which can further provide data augmentation for IR corpora. In the future, it would be interesting to investigate whether unfreezing the context encoder during training would lead to improved retriever performance.

Acknowledgements

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) through the projects CoRA4NLP (01IW20010), XAINES (01IW20005), and IMPRESS (01IS20076). We thank the anonymous reviewers for their constructive feedback.

[7] https://github.com/facebookresearch/DPR.git

References

Petr Baudiš and Jan Šedivỳ. 2015. Modeling of the question answering task in the yodaqa system. In *International Conference of the cross-language evaluation Forum for European languages*, pages 222–228. Springer.

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.

Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–199.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Mandar Joshi, Eunsol Choi, Daniel 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, Vancouver, Canada. Association for Computational Linguistics.

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)*. Association for Computational Linguistics.

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.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Mike Lewis and Angela Fan. 2018. Generative question answering: Learning to answer the whole question. In *International Conference on Learning Representations*.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. *Advances in Neural Information Processing Systems*, 33:18470–18481.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. 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, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020c. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xueguang Ma, Kai Sun, Ronak Pradeep, and Jimmy Lin. 2021. A replication study of dense passage retriever. *arXiv preprint arXiv:2104.05740*.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint*, 6.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797.

Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau, and Manzil Zaheer. 2023. Questions are all you need to train a dense passage retriever. *Transactions of the Association for Computational Linguistics*, 11:600–616.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

Stalin Varanasi, Saadullah Amin, and Günter Neumann. 2021. Autoeqa: Auto-encoding questions for extractive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4706–4712.

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.

Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. Sparta: Efficient open-domain question answering via sparse transformer matching retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 565–575.