Optimizing Short Text Information Retrieval With Dense Passage

Amit Kumar Singh Sanger
Computer Science
KIET Group Of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
amit.sanger@kiet.edu

Rajendra Kumar Patel
Computer Science
KIET Group Of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
rajendra.patel@kiet.edu

Sheersh Sharma
Computer Science
KIET Group Of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
sheersh.2125cs1098@kiet.edu

Aditya Pandey
Computer Science
KIET Group Of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
aditya.pandey.31a@gmail.com

Shivansh Singh
Computer Science
KIET Group Of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
shivansh.2125cs1127@kiet.edu

Aryan Srivastava
Computer Science
KIET Group Of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
aryan.2125cs1102@kiet.edu

Abstract:

DPR (Dense passage retrieval) is one of the famous technique from natural language processing which gives efficient results for a system of question-answering. The de facto method for answering open-domain questions is to select appropriate contexts by employing conventional sparse vector space models, such as BM25 or TF-IDF, for effective passage retrieval. This research demonstrates that retrieval may be achieved effectively using dense representations, where embeddings are picked up from a small amount of passages and questions utilising a simple dual encoder framework. Test results on multiple open-domain QA datasets, with numerous open-domain QA benchmarks are obtained are studied in this research. Multiple state of art techniques and their limitations are discussed in this research paper. This survey paper even includes many researchers study as well as pretrained language models.

Keywords : DPR(Dense Passage Retrieval), Pretrained Language Model , question answering.

I. INTRODUCTION

In the realm of natural language processing (NLP), open-domain question answering (QA) presents a challenging task where the goal is to retrieve precise answers to questions from a vast pool of unstructured textual data. Traditional approaches to this problem often rely on information retrieval (IR) techniques, which may struggle to capture nuanced semantic relationships between questions and answers. However, recent advancements in AI have introduced a ground-breaking method known as Dense Passage Retrieval (DPR), which promises to revolutionise open-domain QA by leveraging dense vector representations of passages.

Factoid questions are addressed via open-domain question answering (QA) by employing an extensive collection of materials. Although early QA systems were frequently complex, with many different components; Moldovan et al. (2003), reading comprehension models have advanced to the point where a much simpler two-stage framework is suggested: the retrieved contexts review carefully by machine reader and determine the right response after: (1) A context retriever initially chooses a small subset of passages, some of which contain the answer to the inquiry (Chen et al., 2017). Even though switching from open-domain QA to machine reading is a perfectly logical strategy, practice 2 frequently shows a significant decrease in performance, highlighting the need for improved retrieval.

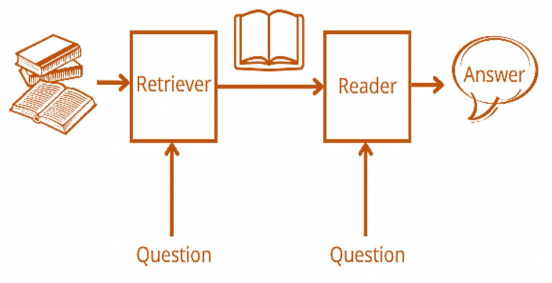

Figure : Pipeline of Question Answering

There are most commonly used techniques includes BM25 and TF-IDF. The limitations of existing techniques is , it doesn’t have capability to retrieve the text if it is not matching exactly. Recently many researchers found that information can be easily and more effectively retrieved using dense vector representation of text.

DPR represents a significant leap forward in the field, offering a more sophisticated approach to matching questions with relevant passages. By encoding both questions and passages into dense vector embeddings, DPR facilitates more accurate and contextually rich retrieval, leading to improved performance in QA tasks. This introduction aims to explore the key concepts and implications of dense passage retrieval for open-domain question-answering.

Additionally, it has two flaws with QA datasets. Firstly, there is a lot of computation involved in ICT pretraining, and it's not quite apparent if regular phrases are a suitable substitute for the questions in the objective function. Second, the relevant representations might not be ideal since the context encoder is not adjusted by question-answer pairings. In this research, we investigate the following question: can we train a more effective dense embedding model without additional pretraining, using simply pairs of questions and passages.

We have a very robust Dense Passage Retriever (DPR). It performs significantly better than BM25 in both top-5 accuracy (65.2% vs. 42.9%) and end-to-end QA accuracy (41.5% vs. 33.3%) when compared to ORQA for the open natural questions. We have two things to offer. First, we show that optimising the question and passage encoders on current question-passage pairings is enough to significantly outperform BM25, provided the right training configuration is used.

Almost all NLP applications were heavily utilising TF-IDF sparse vector algorithms for text-based answers, but these algorithms required heavy training on a large corpus or dataset. To overcome this issue, the author of this paper employs dense-based word embedding techniques to answer questions. This dense-based work embedding can be generated for any text without requiring any further training.

Converting any text into a numeric vector is called dense embedding, and the author utilises a BERT-based algorithm to convert text into a dense numeric vector. This vector will be processed through the inner DOT product to measure similarity between the question vector and predict accurate answers. The proposal algorithm requires no heavy training and can compete with any advanced answering tool like BM2.5. In this research author is comparing Dense Passage Retrieval technique with many other algorithms. Each algorithm performance is evaluated using Accuracy and tested on many datasets.

Training algorithm on all datasets may consume more time so we suggested to use WEB-QUESTION dataset available on KAGGLE repository. Most of the authors used pretrained language model as datasets size is above 5 GB whose training is not possible (take longer time) on normal laptops.

II. LITERATURE SURVEY

On open-domain Question Answering (QA), where latent representations of passages and questions are utilised for maximal inner product search during the retrieval process, dense neural text retrieval has demonstrated encouraging results. Nevertheless, existing dense retrievers heavily depend on the splitting process and necessitate breaking documents into brief sections that typically contain local, incomplete, and occasionally biassed information. Consequently, it could produce misleading and hidden representations, which would lower the quality of the final retrieval outcome. In this work, we propose Dense Hierarchical Retrieval (DHR), a hierarchical framework that may leverage both microscopic semantics unique to each passage and macroscopic semantics in the document to build accurate dense representations of passages. To be more precise, a passage-level retriever retrieves appropriate portions from papers that a document-level retriever has first identified as relevant. By considering the significance of the retrieved sections at the document level, the ranking will be adjusted even further. Furthermore, two negative sampling techniques—In-Doc and In-Sec negatives—as well as a hierarchical title structure are examined. We use extensive open-domain QA datasets and apply DHR to them. DHR provides a major improvement over the original dense passage retriever and aids in surpassing the strong baselines on several open-domain QA benchmarks with an end-to-end QA system. [1]

Dense passage retrieval has emerged as a novel paradigm in open-domain question answering to retrieve pertinent passages for solutions. In order to learn dense representations of questions and passages for semantic matching, the dual-encoder architecture is typically used. The lack of labelled positives, training data shortages, and inconsistency between training and inference present significant hurdles to the effective training of a dual encoder. We provide an enhanced training method, known as RocketQA, to overcome these difficulties and enhance dense passage retrieval. Our three main technical contributions to RocketQA are data augmentation, denoised hard negatives, and cross-batch negatives. According to the experiment results, RocketQA performs significantly better on MSMARCO and Natural Questions than earlier state-of-the-art models. We also carry out comprehensive studies to investigate the performance of the three RocketQA techniques. [2]

In a bi-encoder design dense passage retrievers (DPR) is the one key feature that use of passage encoder and separate question. Prior efforts towards the generalisation of DPR have focused on testing both encoders simultaneously on domain adaptation, or out-of-distribution (OOD) question-answering (QA) tasks. But, the impact of DPR's unique question/passage encoder on generalisation remains unknown. In particular, this paper aims to investigate the generalisation potential of an IND question/passage encoder when combined with an OOD passage/question encoder from a different domain. In order to address this, we examine several combinations of the passage encoder that was taught from five benchmark QA datasets and the DPR's question on both in-domain and out-of-domain questions. It appears that the question encoder generally affects the upper bound of generalisation, but the passage encoder has a greater impact on the bottom bound. Applying an OOD passage encoder, for instance, typically reduces retrieval accuracy, but using an OOD question encoder may even increase accuracy. [3]

Without resorting to outside knowledge, generative models for open-domain question-answering have proven to be competitive. Although promising, this method requires the deployment of expensive training and query models with billions of parameters. In this study, we examine the extent to which these models can be made more useful by obtaining text passages that may include evidence. We achieve cutting-edge outcomes on the open TriviaQA and Natural Questions benchmarks. It's interesting to note that when the number of recovered paragraphs increases, this technique performs noticeably better. This demonstrates how flexible the framework provided by sequence-to-sequence models may be in effectively combining and aggregating evidence from numerous sections. [4]

Dense retrieval has become the principal technique for open-domain question answering (OpenQA) in recent years. Nevertheless, prior studies frequently overlooked the significance of the passage side in favour of the query side. For better OpenQA performance, we believe that the question and passage sides are equally significant and should be taken into account. In this work, we suggest a contrastive pseudo-labeled data set built on questions and passages independently. We utilise a knowledge-filtering approach in conjunction with an enhanced pseudo-relevance feedback (PRF) algorithm to enhance the semantic information within dense representations. Furthermore, to update the dense representations iteratively, we suggested an Auto Text Representation Optimisation Model. The outcomes of our experiments show that our techniques efficiently optimise dense representations, increasing their distinguishability in dense retrieval and enhancing the overall performance of the OpenQA system. [5]

An response to an open-domain issue is provided using data that has been gathered from a large corpus. For training, state-of-the-art neural techniques require annotations of intermediate evidence. Nevertheless, the cost of these intermediary annotations makes approaches that depend on them unsuitable for the more typical condition in which question-answer pairs are the only data available. The purpose of this research is to determine whether models can be trained to find evidence from a large corpus using only remote supervision from answer labels, which would eliminate the need for additional annotation costs. We provide a new method called DISTDR, which iteratively outperforms a weak retriever by alternating between finding evidence from the most recent model and letting the model learn what evidence is most possible. Our research demonstrates that DISTDR finds more precise evidence over iterations, resulting in improved models. [6]

For various open-domain question-answering (QA) tasks, documents from a common corpus (like Wikipedia) can be retrieved using multi-task dense retrieval models. Datasets like Trivia and NQ cover more entries, SQuAD only focuses on a limited number of Wikipedia articles. As a result, joint training on their union may result in performance reduction. In order to address this issue, we suggest training separate dense passage retrievers (DPR) for various tasks and combining their predictions during testing. Uncertainty estimate is then used as weights to indicate the probability that a given research belongs within the authority of each expert. Furthermore, we demonstrate that, when applied to a mixed subset of various QA datasets, our technique outperforms the joint-training DPR in handling corpus inconsistency. [7]

Author [year]	Proposed technique	Limitations
Ye Liu,etAl [2021]	Use of macroscopic as well as micropic semantics used with Dense Hierarchical Retrieval (DHR)	Scalability and complexity may be little high
Qu,	RocketQA: ,	Training

Ding,et Al [2020]	Architecture of dual auto-encoder is used	model may take higher time
Li, Minghan,etAl [2021]	Encoders with question and answer are analysed in proposed method	Varying results , domain to domain
Izacard Gautier [2020]	Multiple passage based evidence aggregation	Many limiting factors may include in proposed model
Zhai , Q, Zhu[2023]	PRF(pseudo-relevance feedback)	Results vary based on labeled pseudo data
Chen Zhao ,etal [2021]	Evidence of learning and finding is used (DISTDR)	Varing quality as per answer labels
Li, Minghan	Multi-task Dense Retrieval	Based on distribution and diversity

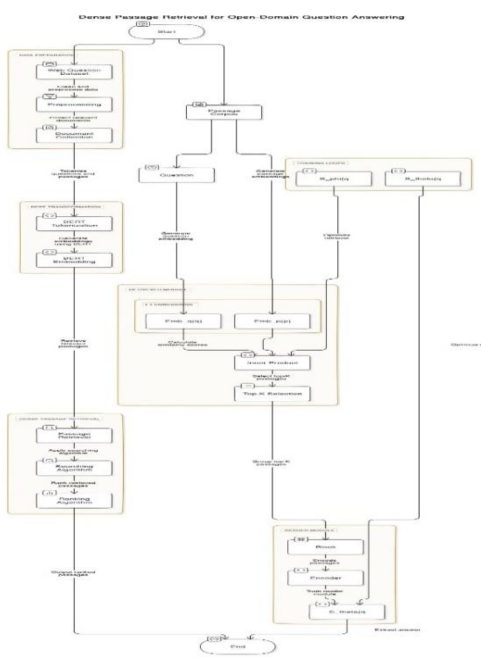

Fig : Systematic Method For Dense Passage Retrieval for Open Domain Question Answering

There are many existing techniques which are used in this field which are as described below –

Aspect-Based Neural Models: Neural models, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers, are employed to create aspect-aware representations for both queries and documents. These models can capture nuanced aspects of the input data.

Aspect Fusion Strategies: Fusion strategies like late fusion, early fusion, or cross-modal fusion are used to integrate aspect-specific information effectively, combining representations from different aspects to generate a comprehensive understanding of the query or document.

Graph Neural Networks (GNNs): GNNs are utilized to model relationships and interactions between aspects in a query or document. They enable capturing complex cross-aspect interactions and are effective in incorporating structural information within the aspect representation.

Reinforcement Learning: Reinforcement learning approaches are employed to optimize the ranking of documents based on multiple aspects. These methods use rewards to guide the model in selecting documents that align well with the various aspects of the query.

Aspect-Attention Networks: Models with specialized attention mechanisms, such as aspect-level attention, are designed to highlight relevant aspects within the input. These mechanisms help in effectively utilizing aspect-specific information during retrieval.

Proposed Techniques

Dual Encoder Models: Neural models, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers, are employed to create aspect-aware representations for both queries and documents. These models can capture nuanced aspects of the input data

Cross-Encoder Models: Cross-encoder models jointly encode query-document pairs, capturing rich interactions for improved accuracy. However, their higher computational cost makes them less suitable for large-scale or real-time retrieval compared to dual-encoder models.

Approximate Nearest Neighbor (ANN) Search: ANN search algorithms enable efficient large-scale retrieval by approximating nearest-neighbor searches in high-dimensional spaces, allowing quick identification of relevant records within massive databases.

Knowledge Distillation: Knowledge distillation transfers knowledge from a large, complex model (the teacher) to a smaller, efficient one. This enables deploying DIR models in resource-constrained environments without significantly compromising performance or speed.

Proposed work

In Proposed work used compute loss function, cross entropy hard negative and positive function as well as SoftMax for which increase accuracy of MRR , Exact match and Recall in comparison of base paper.

Exact Match EM: It reflects the model's precision in ranking the most relevant document first. Our model's EM steadily improves, reaching 0.985 in the final epoch, surpassing Rocket QA's 0.957
Rocket QA paper: 0.957
Current Model: 0.985
Initial Epoch (Epoch 1): 0.737
Final Epoch (Epoch 10): 0.985

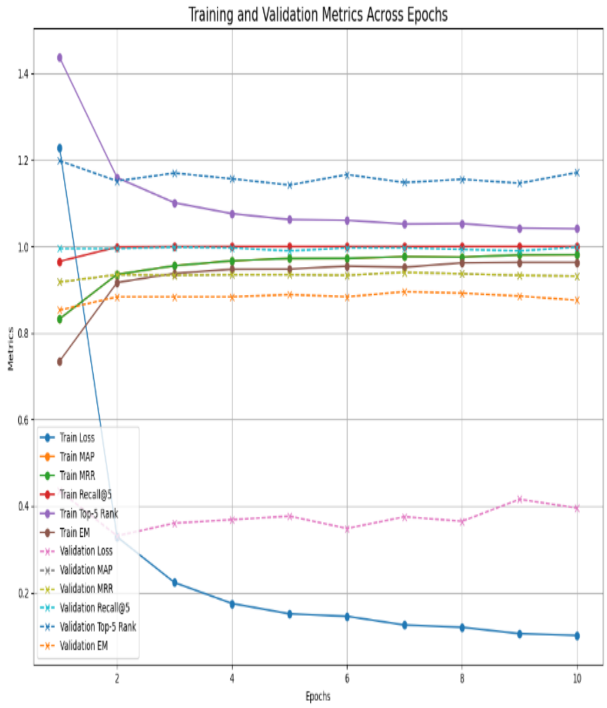

Fig : Training And Validation Metrics Across Epochs

S.No.	Paper and author	Recall @20	EM
1	Ye Liu, Kazuma Hashimoto, Ying bo Zhou, Semih Yavuz, Caiming Xiong, Philip S. Yu "Dense Hierarchical Retrieval for Open-Domain	**91.3**	**41.5**

	Question Answering”		
2	Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., ... & Wang, H. (2020). RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.	**95.3**	**56**
3	Li, Minghan, and Jimmy Lin. "Encoder adaptation of dense passage retrieval for open-domain question answering." *arXiv preprint arXiv:2110.01599* (2021).	**92.3**	**48.5**
4	Izacard, Gautier, and Edouard Grave. "Leveraging passage retrieval with generative models for open domain question answering." *arXiv preprint arXiv:2007.01282* (2020).	**88.7**	**47**
5	Zhai, Q.; Zhu, W.; Zhang, X.; Liu, C. Contrastive Refinement for Dense Retrieval Inference in the Open-Domain Question Answering Task. *Future Internet* **2023**, *15*, 137.	**91.7**	**52.1**
6	Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. 2021. Distantly-Supervised Dense Retrieval Enables Open-Domain Question Answering without Evidence Annotation. In *Proceedings of the 2021 Conference on Empirical*	**92.0**	**49.2**

	Methods in Natural Language Processing, pages 9612–9622, Online and Punta Cana, Dominican Republic. Association.		
7	Li, Minghan, et al. "Multi-task dense retrieval via model uncertainty fusion for open-domain question answering." *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021.	**93.2**	**53.5**
8	**Optimizing Short Text Information Retrieval With Dense Passage**	**1.000**	**98.5**

Conclusion

In this research many previous author works are studied and even their limitations are studied to understand the details of question answering. In this research limitations of efficiency and cost effectiveness are found common. Most common technique used found are pretrained language models (PLMs) as well as BM25 and TF-IDF. Using this previous authors study, new open domain question answering is suggested to implement which provides more efficiency, cost effective and accuracy for QA.

In this paper work we used compute loss function, cross entropy hard negative and positive function as well as SoftMax for which increase accuracy of MRR , Exact match and Recall in comparison of base paper is found and by combining these compute functions we can get better output in dense passage retrieval and can be used in future with some modification in algorithms to get better and efficient output.

REFERENCES

1. Ye Liu, Kazuma Hashimoto, Yingbo Zhou, Semih Yavuz, Caiming Xiong, Philip S. Yu "Dense Hierarchical Retrieval for Open-Domain Question Answering" https://doi.org/10.48550/arXiv.2110.15439.
2. Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., ... & Wang, H. (2020). RocketQA: An

optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

3. Li, Minghan, and Jimmy Lin. "Encoder adaptation of dense passage retrieval for open-domain question answering." *arXiv preprint arXiv:2110.01599* (2021).
4. Izacard, Gautier, and Edouard Grave. "Leveraging passage retrieval with generative models for open domain question answering." *arXiv preprint arXiv:2007.01282* (2020).
5. Zhai, Q.; Zhu, W.; Zhang, X.; Liu, C. Contrastive Refinement for Dense Retrieval Inference in the Open-Domain Question Answering Task. *Future Internet* **2023**, *15*, 137. https://doi.org/10.3390/fi15040137
6. Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. 2021. Distantly-Supervised Dense Retrieval Enables Open-Domain Question Answering without Evidence Annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9622, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
7. Li, Minghan, et al. "Multi-task dense retrieval via model uncertainty fusion for open-domain question answering." *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021.
8. https://towardsdatascience.com/understanding-dense-passage-retrieval-dpr-system-bce5aee4fd40