# Diverse Multi-Answer Retrieval with Determinantal Point Processes

**Poojitha Nandigam**[*]**, Nikhil Rayaprolu**[*]**, Manish Shrivastava**
Language Technologies Research Centre (LTRC)
International Institute of Information Technology, Hyderabad, India
poojitha.nandigam@research.iiit.ac.in, nikhil.rayaprolu@students.iiit.ac.in,
m.shrivastava@iiit.ac.in

## Abstract

Often questions provided to open-domain question answering systems are ambiguous. Traditional QA systems that provide a single answer are incapable of answering ambiguous questions since the question may be interpreted in several ways and may have multiple distinct answers. In this paper, we address multi-answer retrieval which entails retrieving passages that can capture majority of the diverse answers to the question. We propose a re-ranking based approach using Determinantal point processes utilizing BERT as kernels. Our method jointly considers query-passage relevance and passage-passage correlation to retrieve passages that are both query-relevant and diverse. Results demonstrate that our re-ranking technique outperforms state-of-the-art method on the AmbigQA dataset.

## 1 Introduction

The objective of open-domain question answering is to provide answers to queries utilising a large collection of documents from the World Wide Web, Wikipedia etc. More than 50% of questions present in a widely used open-domain QA dataset (Natural Questions Kwiatkowski et al. (2019)) comprise of ambiguous questions (Min et al. (2020)). Questions that are ambiguous may be interpreted in a number of ways and as a result, they need various answers. In this paper, we concentrate on questions with multiple distinct answers.

Open domain question-answering systems are designed to generate answers from several data sources. Since similar information can be present across multiple data sources, it introduces a significant amount of redundancy. Traditional open-domain QA (Chen et al. (2017)) systems comprise of a Retriever, which retrieves passages relevant to the question. A passage retriever is primarily concerned with retrieving passages that are relevant

to the query, and it does not address redundancy in the passages during retrieval. To be able to produce diverse answers to the question, the passages retrieved must be both relevant to the question and distinct from one another. After the retrieval stage, we introduce a novel re-ranking approach to handle redundant passages. As a result, the re-ranked passages would capture most of the diverse answers to the question. In this paper, we investigate the multi-answer retrieval task, which entails retrieving passages that can cover the distinct answers.

Re-ranking methods have been employed previously to improve the question answering accuracy significantly(Wang et al. (2019);Nogueira and Cho (2019); Min et al. (2021); Clark and Gardner (2017)). Min et al. (2021) tackles diverse multi-answer retrieval by proposing a re-ranker based on an auto-regressive framework in which each passage selected is dependent on the passages chosen at a previous time step.

Determinantal Point Processes (DPP) (Kulesza and Taskar (2012)) are probabilistic models that are effective at identifying diverse subsets of elements from a collection while preserving quality. DPP methods have proven effective in natural language processing tasks where there is a need for diverseness. Cho et al. (2019), Li et al. (2019), and Cho et al. (2020), Sharghi et al. (2017) have used DPPs to perform summarization by choosing salient but also diverse items to be included in the summaries. In this paper, we propose an unsupervised re-ranking technique for multi-answer retrieval utilising Determinantal point processes and BERT to model the kernels.

Our contributions can be summarized as follows:
1) We propose a re-ranking method based on determinantal point processes that focuses on diverse passage retrieval.
2) Since our approach is unsupervised, our method does not require a large amount of data unlike prior re-ranking methods(Min et al., 2021). Instead, we

---

[*]Equal contribution

rely on DPP to identify the most relevant passages to the question that are distinct from one another. 3) We demonstrate that our technique outperforms the state-of-the-art method on the AmbigQA dataset using MRECALL @ $k$ metrics.

## 2 Related Work

Many open domain question answering systems(Chen et al. (2017); Yang et al. (2019); Izacard and Grave (2021); Guu et al. (2020); Lee et al. (2019)) adopt the *retriever-reader* method by retrieving the relevant documents and later applying neural techniques to predict the answer. The *retriever-reader* method was first proposed by Chen et al. (2017). DrQA(Chen et al. (2017)) uses Wikipedia as knowledge source and employs a sparse retrieval method using TF-IDF and a recurrent neural network to identify the answer spans. While Yang et al. (2019) adopts Anserini retriever(Yang et al. (2017)) using BM25 as the ranking function and BERT model (Devlin et al. (2018)) as the reader. Sparse retrieval based methods, such as TF-IDF and BM25, face challenges when retrieving relevant passages that do not match the question's exact terms. Dense retrieval-based approaches, on the other hand, overcome this problem by mapping each word into a vector space in which words with similar meanings tend to be closer together. ORQA (Lee et al. (2019)) and DPR (Karpukhin et al. (2020)) employ a question and passage encoder based on BERT and compute a relevance score. Using this relevance score, the retriever retrieves the most relevant documents from the corpus.

## 3 Determinantal Point Processes for Re-ranking

Re-ranker acts as a filter to pick a limited number of passages that can be used as input to generate answers to the questions. We formulate the task of passage re-ranking as a subset selection problem. Our objective is to choose a subset of passages ($Y$) of size $k$ from the ground set $\mathcal{Y}$ comprising N passages that covers all of the answers to a given question $q$. DPP models a distribution on all the subsets of the ground set $\mathcal{Y}$ jointly considering the quality and diversity. A subset $Y$ is drawn according to the probability distribution $P$.

$$P(Y; L) \propto det(L_Y) \qquad (1)$$

$$P(Y; L) = \frac{\det(L_Y)}{\det(L + I)} \qquad (2)$$

where $I$ is the identity matrix, $L \in \mathbb{R}^{N \times N}$ is a positive semi-definite matrix referred as *L-ensemble*, $\det(.)$ denotes the determinant of a matrix, and $L_y$ is the submatrix of $L$ indexed by items in $Y$. $L$ matrix jointly considers query-passage relevance as well as passage-passage correlation through eq. 3.

$$L_{ij} = Q(i, q) \cdot S(i, j) \cdot Q(j, q) \qquad (3)$$

DPP focuses on two measures - quality and similarity ( Fig 1). Quality score $Q(i, q)$ measures how salient the passage $i$ is and whether it contains an answer to the question $q$. Similarity score $S(i, j)$ is computed between two passages $i$ and $j$ to incorporate diversity in the passages. DPP assigns a probability to a set Y proportional to the determinant of *L-ensemble* which may be interpreted geometrically as the volume of the parallelepiped covered by the quality and similarity measures (Kulesza and Taskar (2012)). A diverse passage subset occupies more volume than a subset of similar passages, therefore DPP assigns higher probability to diverse and relevant passages rather than the most relevant and similar passages. If passages are relevant and diverse, then the passages can cover multiple distinct answers to the question.
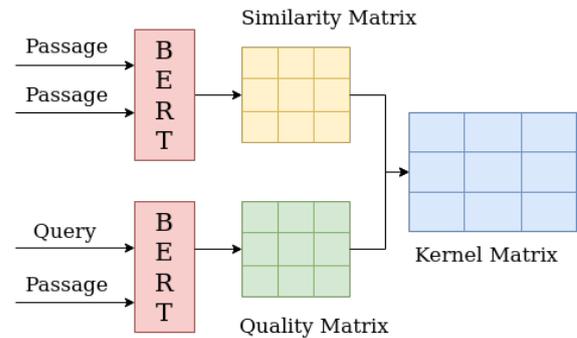


Figure 1: An overview of the proposed re-ranking method using DPP. A similarity score between the passages and a quality score between the question and passage are computed. These two scores are utilised to construct the DPP kernel matrix.

### 3.1 BERT for Similarity matrix

To compute the similarity scores, we use a pretrained BERT model (Devlin et al. (2018); Reimers and Gurevych (2019)) to generate embeddings for every passage. The model takes the passage as

input and produces a 768 dimensional dense embedding. We use these embeddings to calculate the cosine similarity of all passages and compute a similarity matrix $S \in \mathbb{R}^{N \times N}$ for the whole passage set. All the values in the similarity matrix lie in the range of $\lceil 0, 1 \rceil$. If passages $i$ and $j$ are similar, the similarity value $S(i, j)$ lies closer to 1, if they are distinct, the value lies closer to 0, and if $i = j$, $S(i, j)$ becomes equal to 1.

$$S(i, j) = cosine\_sim(BERT_A(i), BERT_A(j)) \tag{4}$$

### 3.2 BERT for Quality matrix

We use a pretrained BERT model trained on MS MARCO (Nguyen et al. (2016)) for computing the Quality matrix. The model takes in a query and a passage and generates the quality score. Higher quality score indicates that the passage is most relevant to the query and therefore most likely to answer the query. Unlike for computing similarity matrix, we do not perform cosine similarity over the model's outputs to produce a score, instead, we use a BERT encoder that concatenates both query and passage and generates a score. The quality matrix $Q \in \mathbb{R}^{N \times N}$ is computed by performing the matrix multiplication of the scores ($N \times 1$) with it's transpose resulting in $N \times N$ dimensioned vector . These quality scores are then normalized to lie between $\lceil 0, 1 \rceil$.

$$Q(i, j) = Norm(BERT_B([i; j])) \tag{5}$$

### 3.3 Sampling

Traditional DPP sampling algorithms have higher run-time complexity when $L$ matrix is large. We apply an efficient sampling technique - BFGMInference (Li et al. (2019);Chen et al. (2018)). BFGMInference approximates a greedy approach to select a passage that maximizes the $det(L_Y)$ and adds it to the passage subset.

$$f(Y) = \log \det (L_Y)$$
$$k = \underset{i \in \mathcal{Y} \setminus Y}{\arg \max} f(Y \cup \{i\}) - f(Y) \tag{6}$$

## 4 Experiments

In this section, we discuss about the passage retrieval method, the dataset we used in our experiments, the evaluation metric, and the results of our experiments.

### 4.1 Passage retrieval

Wikipedia is utilised as the corpus for retrieving passages for the questions. Each Wikipedia article is broken into multiple passages containing the same number of words. We retrieve query-relevant passages from Wikipedia using the Dense Passage Retriever (DPR) (Karpukhin et al. (2020); Lin et al. (2021)). DPR computes encodings for all passages extracted from the Wikipedia corpus and builds an index. The inner product of the query and passage encodings is used to determine the similarity scores between them. Passages with the highest scores are the ones that are most relevant to the query, and these passages are subsequently sent into the re-ranker as input.

### 4.2 Dataset

We evaluated our method on an open-domain question-answering dataset AmbigQA (Min et al. (2020)), which contains multiple-answer questions. The dataset was created from an anonymised collection of Google search queries submitted by users seeking information on different subjects. It consists of 14,042 question-answer pairs derived from the Natural questions dataset (Kwiatkowski et al. (2019)) and is split into train, validation, and test sets. Train set consists of 10,036 question-answer pairs, validation set consists of 2,002 examples, and test set consists of 4,042 examples.

### 4.3 Evaluation metric

MRECALL @ $k$ (Min et al. (2021)) is used to evaluate the re-ranking of passages for questions with diverse answers. As per this metric, if a query has $n$ answers, the $k$ passages that are retrieved must cover all of the answers. If $n <= k$, all answers must be covered; if $n > k$, the passages retrieved must contain at least $k$ answers. A retrieval is deemed successful if the passages retrieved include all or at least $k$ of the answers to the query.

### 4.4 Results

We compare our technique to a few additional baselines, all of which were assessed using the MRECALL @ $k$ metric on the AmbigQA dataset.

- **DPR$^+$** Min et al. (2021) integrates REALM (Guu et al. (2020)) with DPR (Karpukhin et al. (2020)). As described in Section 4.1, DPR is a dense retrieval based technique that utilizes the FAISS library to retrieve the relevant documents. Encoders for the query and passage

| Models | Top5 | Top 10 |
|---|---|---|
| | AmbigQA-Dev | AmbigQA-Dev |
| DPR$^+$(Min et al. (2021)) | 55.2/36.3 | 59.3/39.6 |
| DPR$^+$ + Nogueira and Cho (2019) | 63.4/43.1 | 65.8/46.4 |
| JPR(Min et al. (2021)) | 64.8/45.2 | 67.1/48.2 |
| QRR | 62.0/42.3 | 70.8/57.6 |
| DPP-R | **66.9/53.5** | **72.8/58.8** |

Table 1: Performance of various models on AmbigQA dataset. Each row contains the MRECALL @ $k$ metrics for single answer retrieval and multi-answer retrieval respectively.

are initialized using REALM and the DPR training method is followed.

- **DPR$^+$ + Nogueira and Cho (2019)** employs DPR$^+$ for the first stage of retrieval and the re-ranking method in Nogueira and Cho (2019) is applied on the retrieved passages.

- **JPR** Min et al. (2021) employs DPR$^+$ as the initial ranker and an auto-regressive framework is adopted as a re-ranker to generate diverse passages.

- **Query Relevance Re-ranking(QRR)** In this method, we first calculate the quality scores for each passage (described in section 3.2) and then we sort the passages based on these scores to pick the top-$k$ passages. Here, similarity among the passages is not considered.

- **DPP-R** We employ our method described in section 3 to retrieve highly diverse and relevant passages.

We calculate the performance of diverse multi passage retrieval using the MRECALL @ $k$ measure described in section 4.3. Evaluation on the AmbigQA dataset demonstrates that our approach outperforms existing re-ranking techniques. Our technique requires no human annotations for multi passage re-ranking while outperforming existing methods, as shown in Table 1. DPP is modelled to select a subset of high-quality and diverse passages, which contributes to the success of our method for this task. Experiments demonstrate that the DPP-based technique achieves promising results for retrieving passages containing diverse answers.

## 5 Discussion

**Impact on QA system's performance:** An Open domain question answering system's pipeline consists of three stages. 1) Retrieval 2) Re-ranking followed by 3) Answer extraction. Improvements in

any of these stages significantly improve the overall system's ability to answer a question. Nogueira and Cho (2019), Min et al. (2021) have shown that the use of a re-ranker has led to end-to-end QA improvements. Based on the results presented in Table 1, the DPP method enhances re-ranking for both single and multi-answer questions. We believe that this improvement in re-ranking will also improve the overall performance of the end-to-end QA system.

**Impact of diversity:** DPP-R and JPR retrieve diverse passages utilising DPP and auto-regressive framework, respectively. Other approaches like QRR, retrieve just passages that are relevant to the query and do not tackle passage redundancy. We observe that our approach using DPP performs better than the QRR method. In order to re-rank, QRR simply considers how relevant a passage is to the query, and it retrieves the top-k passages with the highest relevance score for a given query. On the other hand, DPP-R takes into account how relevant the passage is to the query and also how similar passages are to each other, in order to eliminate redundant passages leading to diversity in the retrieved passages. DPP-R and JPR outperform other methods that do not emphasise diversity in multi-answer retrieval. For single answer retrieval, DPP-R and JPR have fared better than other methods, with the minor exception that QRR beats JPR in top-10 re-ranking. This demonstrates that diversity is an important aspect to consider during the re-ranking stage.

## 6 Conclusion

In this paper, we propose a DPP-based approach to improve the diverseness of the retrieved passages. We compare our method to the state-of-the-art method and outperform it by 3% (top 5), 8% (top 10) for single-answer questions, and 18% (top5) and 21% (top10) for multi-answer retrieval

on AmbigQA dataset.

## References

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

Laming Chen, Guoxin Zhang, and Eric Zhou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in Neural Information Processing Systems*, 31.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2020. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1027–1038.

Sangwoo Cho, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2019. Multi-document summarization with determinantal point processes and contextualized representations. *arXiv*.

Christopher Clark and Matt Gardner. 2017. Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming Wei Chang. 2020. REALM: Retrieval-Augmented language model pre-training. In *37th International Conference on Machine Learning, ICML 2020*, volume PartF16814, pages 3887–3896.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 874–880.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering.

Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Lei Li, Wei Liu, Marina Litvak, Natalia Vanetik, and Zuying Huang. 2019. In conclusion not repetition: Comprehensive abstractive summarization with diversified attention based on determinantal point processes. *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*, pages 822–832.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations. *arXiv preprint arXiv:2102.10073*.

Sewon Min, Kenton Lee, Ming-Wei Chang, Kristina Toutanova, and Hannaneh Hajishirzi. 2021. Joint passage ranking for diverse multi-answer retrieval. *arXiv preprint arXiv:2104.08445*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Aidean Sharghi, Jacob S Laurel, and Boqing Gong. 2017. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4788–4797.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1256.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with BERTserini. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Demonstrations Session*, pages 72–77.