Topic-Mono-BERT: A Joint Retrieval-Clustering System for Retrieving Overview Passages

Sumanta Kashyapi
University of New Hampshire
Durham, USA
Sumanta.Kashyapi@unh.edu

Laura Dietz
University of New Hampshire
Durham, USA
dietz@cs.unh.edu

ABSTRACT

For most queries, the set of relevant documents spans multiple subtopics. Inspired by the neural ranking models and query-specific neural clustering models, we develop Topic-Mono-BERT which performs both tasks jointly. Based on text embeddings of BERT, our model learns a shared embedding that is optimized for both tasks. The clustering hypothesis would suggest that embeddings which place topically similar text in close proximity will also perform better on ranking tasks. Our model is trained with the Wikimarks approach to obtain training signals for relevance and subtopics on the same queries.

Our task is to identify overview passages that can be used to construct a succinct answer to the query. Our empirical evaluation on two publicly available passage retrieval datasets suggests that including the clustering supervision in the ranking model leads to about 16% improvement in identifying text passages that summarize different subtopics within a query.

ACM Reference Format:
Sumanta Kashyapi and Laura Dietz. 2022. Topic-Mono-BERT: A Joint Retrieval-Clustering System for Retrieving Overview Passages. In *Forum for Information Retrieval Evaluation (FIRE '22), December 9–13, 2022, Kolkata, India.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3574318.3574336

1 INTRODUCTION

Even unambiguous search queries are expected to have different subtopic dimensions. As a result, the retrieved text rankings contain multiple relevant subtopics; containing some passages that are more specific to individual topics, while other passages are more general and lend themselves for providing an overview over the query.

For example, let us consider the query: *COVID-19*. Relevant documents for this query may contain information about multiple aspects of the disease such as *symptoms, preventive measures, vaccinations, government policies* and many more. Hence, to gain a holistic knowledge about such a query, the user has to inspect search snippets for each of these links. This is frustrating when the display size is limited, leading to users who abandon their search. Recent work on *"good abandonment"* studies to which extent the user's information can be satisfied with information displayed on the search engine result page (SERP) itself without any further navigation to the web pages [10, 18, 19]. We study the task of ranking passages by how useful they are in providing an overview for the query. This task can be applied to any document ranking to place meaningful information on result pages.

The vision of the TREC CAR track [3] is to answer such queries with a synthetic Wikipedia-like article and has led to the development of different large-scale test collections to study this task [2, 3]. However the TREC CAR passage retrieval tasks only focus on the retrieval step of the process [12, 14, 16]. The next step towards this goal is to distill the information contained in the retrieved passages. In this work, we focus on the distillation step in selecting passages to provide an overview of the query's topic.

Overview passage ranking task: Given a broad query q and a set of relevant text passages P_q (e.g., from retrieved documents), the task is to identify passages $p_i^q \in P_q$ based on their suitability for providing an overview of the query.

Our hypothesis is that for a given broad query q, the knowledge of the relevant subtopics S_q would be helpful for solving this task. For example, the retrieval model could promote passages as central as possible to all subtopics, while demoting passages specific to only a single subtopic.

However, such relevant subtopics S_q are usually not available, and hence need to be clustered. Previous work on using unsupervised clustering and topic modelling in retrieval has only produced mixed results [20]. Here we build on recent work on query-specific subtopic clustering [8] which is supervised with a large benchmark that covers many thematic areas.

To identify suitable overview passages from a given candidate set, we propose Topic-Mono-BERT: a neural BERT-based retrieval system augmented by a subtopic clustering system. While the neural retrieval model will affect the text embeddings to obtain best query-document relevance, the clustering model will influence embeddings of texts on similar subtopics to be in close proximity. Together, the joint system will ensure that the overview passages are highly relevant to the query terms while being topically distinct from all subtopics relevant to the query.

Specifically, we focus on the following research questions:

- **RQ1:** Does the overview passage ranking task benefit from incorporating relevant subtopics? (Answer: yes)
- **RQ2:** Is it important to incorporate *multiple* subtopics as opposed to only considering a relevant and a non-relevant subset? (Answer: yes, multiple)

FIRE '22, December 9–13, 2022, Kolkata, India

ACM ISBN 979-8-4007-0023-1/22/12…$15.00
https://doi.org/10.1145/3574318.3574336

Contributions.

(1) We provide the joint retrieval-clustering model Topic-Mono-BERT with model specification and code online.[1]
(2) We demonstrate that incorporating multiple topics into the ranking criterion helps to perform well on the text ranking task.
(3) We empirically demonstrate that our method outperforms monoBERT, a strong neural retrieval model, by about 16% in terms of mean average precision (MAP).

2 RELATED WORK

The notion of document relevance with respect to a query is closely related to the notion of similarity. If a set of documents are relevant for the same query, it is highly likely that they all discuss similar topics, leading to high inter-document similarity. Document clustering, on the other hand, keeps similar documents on the same cluster while different documents are far apart, essentially relying on a similarity metric. This duality of retrieval and clustering is first explored by Jardine and Rijsbergen [6]. They observe that documents from the same cluster tend to be relevant for similar queries. This is referred to as the *cluster hypothesis*. This suggests that ranking and clustering provide complementary information about a document collection which is often exploited in IR research to improve retrieval models with the help of clustering methods and vice versa. Ailon et al. [1] develop methods to aggregate contradictory information from ranking and clustering. Their methods are applied to ranking and clustering tasks individually and not across the two. Kurland [9] uses query-specific clusters to rerank top k documents of the initial ranking to improve the overall ranking score. However, there is no means of communication between the clustering and retrieval models such that one can take advantage of the other and mutually compensate mistakes. Also, the clustering method is not trainable, so it is only used to provide unsupervised clustering information to the ranking model. He et al. [4] develop an approach to diversify the rankings through clustering. Again this is a forward-only process, meaning there is no feedback from the results to rectify errors in the clustering and ranking system. Liu et al. [13] show how cluster-based retrieval can be beneficial for ranking systems that leverage language models. Instead of using individual documents to model the query-likelihood, the authors first cluster the document set and use the clusters to represent the language model. Tombros et al. [17] document the effects of hierarchical clustering on retrieval results. It has exhaustive related works and experiments supporting the cluster hypothesis by Rijsbergen. Even in the geoscience domain, researchers have found evidence that ranking can be beneficial to clustering [7].

Recent advances in deep learning techniques led to neural retrieval models [5, 15] that achieve better precision than their non-neural counterparts. Moreover, it is possible to easily fine-tune these models for different domains and query-styles. In their multi-stage passage retrieval framework, Nogueira et al. [15] propose monoBERT, a pointwise re-ranking model. We describe this model in detail as our approach builds on it.

[1]https://github.com/nihilistsumo/ORCA

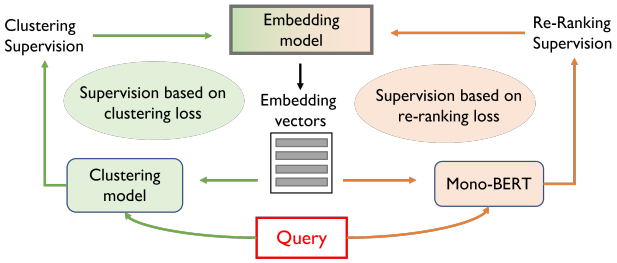


Figure 1: Overall joint-learning architecture of the model where a common embedding model is trained for the Mono-BERT re-ranking model with query-specific clustering supervision.

MonoBERT. The core of the model is a supervised neural reranking model which ranks the passages from the candidate set (in their work produced by a BM25 retrieval). A BERT encoder is trained to emit the relevance scores for each of the query-document pairs. The approach takes advantage of BERT's "next sentence prediction" component, which is retrofit so that for a given query text q, the text of the passage p_i is a valid "next sentence" if and only if the document is relevant. The advantage of this model is that it can leverage cross-attention between query terms and passage terms during the relevance prediction. A suitable joint representation of the query-passage pair can be taken from the the vector associated with the CLS token (denoted $\overrightarrow{p_i^q}$). A final output layer using a Multi-Layer-Perceptron ϕ is trained to generate the relevance score $Rel(q, p_i)$.

$$Rel(q, p_i) = \phi(\overrightarrow{p_i^q}) \tag{1}$$

Despite strong performance of non-neural retrieval models, it is yet to be explored whether augmenting clustering supervision can lead to even better neural retrieval models.

3 APPROACH: TOPIC-MONO-BERT

Building on the success of the neural retrieval models such as monoBERT [11, 15], we explore to which extent integrating them into joint retrieval-clustering system offers further advantages for text ranking tasks.

Neural retrieval models focus only on predicting relevance scores of each query-document pair from the candidate set of documents. As a result, they are not considering the topical diversity of the candidate set. However, for the example task of ranking passages for being suitable to cover relevant subtopics, being informed by subtopics is clearly helpful: those passages are expected to be found in the center of the vector set.

In contrast to passively exploiting topics, in our model we take a more active approach by affecting the underlying embedding space to also represent relevant subtopics. We aim to project query-passage pairs representations $\overrightarrow{p_i^q}$ so that they are representing both (1) relevance for our retrieval task (here: representing an overview for the query), while (2) describing topics through their manifold. In particular, for any two passages p_i, p_j from the same topic, we want their respective query-passage representations to be close—while we want them far apart for passages from different topics.

The approach is depicted in Figure 1. On the right side of the figure, we depict how the training loop of the monoBERT retrieval model (denoted in red) affects the query-passage representations. The left side depicts the supervision from a subtopic clustering module (denoted in green). Our training benchmark provides for every query a ground truth for passage ranking and a ground truth for relevant subtopics.

Instead of directly obtaining the relevance scores from the BERT encoder in the case of monoBERT, in our approach the embedding parameters (referred to as θ from here on) are shared with both the retrieval and clustering module. A subsequent MLP layer ϕ obtains the final relevance score from the embedding model.

Training and model. Our approach trains an embedding model that receives supervision from a subtopic clustering module in addition to a retrieval module by optimizing a joint objective $\mathcal{L}_{rc}$.

$$\mathcal{L}_{rc} = \lambda \cdot \mathcal{L}_r + (1 - \lambda) \cdot \mathcal{L}_c$$

where the retrieval loss function $\mathcal{L}_r$ and clustering loss $\mathcal{L}_c$ are interpolated with scalar calibration parameter λ to adjust the importance of one objective over the other.

For retrieval loss $\mathcal{L}_r$ we use monoBERT's loss function. This allows us to study the benefit of our approach in comparison to monoBERT as a baseline.

For the clustering loss $\mathcal{L}_c$, we optimize the reconstruction of the true adjacency matrix $\mathbf{T}$ of passages versus the predicted adjacency matrix $\mathbf{A}$, with

$$\mathcal{L}_c = \sum_{ij} \left| A_{ij} - T_{ij} \right|$$

Entries of this matrix T_{ij} are set to 1 if passages p_i and p_j are in the same subtopics, while set to 0 if in different subtopics. We obtain a ground truth of the adjacency matrix $\mathbf{T}$ from our training benchmark (detailed in the evaluation).

We use the embeddings to predict an adjacency matrix $\mathbf{A}$ of passages. Where the prediction of A_{ij}, i.e., whether passages p_i and p_j are in the same subtopic, is based on the normalized similarities between passage embeddings $\overrightarrow{p_i^q}$ and $\overrightarrow{p_j^q}$:

$$A_{ij} = \frac{2}{1 + \exp\left\{ \text{sim}\left(\overrightarrow{p_i^q}, \overrightarrow{p_j^q} \right) \right\}}$$

where `sim` is defined to be the similarity function between embedding pairs; specifically, we use the Euclidean distance between the vectors, as used in K-means. We use the logistic form to obtain a matrix entries A_{ij} ranging between 0 and 1, obtaining a smooth and differentiable loss function.

Training benchmark. For training our approach needs to be given a query q along with a set of relevant passages P_q (e.g., retrieved from a BM25 [15], or provided along with the benchmark).

For training, we additionally require information on:

- Relevance: A ground truth of which passages are relevant for retrieval loss $\mathcal{L}_r$.
- Topic clustering: A ground truth adjacency matrix $\mathbf{T}$ indicating which passages are in the same subtopic for the clustering loss $\mathcal{L}_c$.

4 APPLICATION: OVERVIEW RETRIEVER WITH CLUSTERING AUGMENTATION

We use our Topic-Mono-BERT model to develop a system for the task of ranking passages based on their suitability of providing an overview for the query. We refer to this system as **ORCA** (**O**verview **R**etriever with **C**lustering **A**ugmentation).

In this task, overview passages are defined as relevant and we also use them as one ground truth topic cluster. Additionally, the non-overview passages are represented as multiple additional subtopic clusters (although these are all not relevant according to the relevance ground truth for our task). This model would encourage relevant passages to be closer to one another than non-relevant passages, while maximizing the margin between relevant and non-relevant passages. It would also encourage each subtopic to form cohesive clusters.

We are arguing that it is critical to represent multiple subtopics (even if these are not relevant) to influence the embedding space in a beneficial way.

To make this point we are further exploring the following variation, we call *binary clustering*. Instead of having multiple subtopic clusters, which all are negatives with respect to our retrieval task, we only consider two clusters: One cluster of all relevant passages, and one cluster of all negative passages. The difference is that all negatives are encouraged to be close to one another.

In contrast, in our proposed multiple topics variation, only passages within each topic are encouraged to be in close proximity, where passages from different clusters should be far apart.

Regarding **RQ1**, we will study whether incorporating subtopic information will lead to an improvement over a pure ranking model like monoBERT. For **RQ2** we compare the multi-topic version to the binary cluster version to demonstrate that incorporating multiple topics is important.

5 EXPERIMENTAL RESULTS

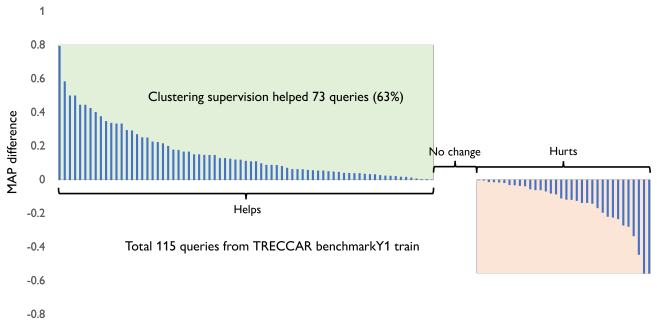
We use our Topic-Mono-BERT model to develop the ORCA system for the task of retrieving passages that constitute good overviews for the given query. We will demonstrate that our model leads to significant improvements over a pure state-of-the-art ranking model (RQ1). We will also show that for obtaining best results, it is important to include multiple subtopics (RQ2).

5.1 Benchmark for overview retrieval

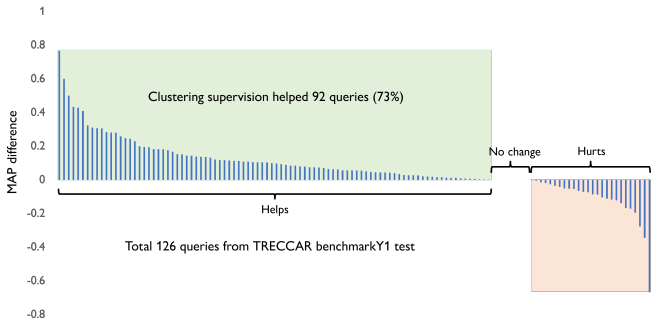
To study the ideal system suitable for our task, we derive a Wikipedia-based retrieval benchmark as depicted in Figure 3. We follow the Wikimarks automatic benchmark creation approach [2], which uses manually selected Wikipedia pages that represent useful information needs.

For each Wikipedia article included in our benchmark, we take the title as the query and construct a candidate set of passages from the paragraphs appearing on the article. Such a candidate set could also be retrieved by some high-quality passage retrieval system, but using this approach allows to compare different re-ranking systems.

On each Wikipedia article the set of overview passages are identified as the passages that appear before the first section (also called



(a) TREC CAR benchmarkY1train queries (despite the name, these are not used for training)



(b) TREC CAR benchmarkY1test queries.

Figure 2: Query-wise analysis for TREC CAR queries. The height of the bars indicates the improvement of MAP score of ORCA over monoBERT.

lead text). These passages are considered relevant for the overview ranking task, and hence are the ground truth for relevance.

For training only, we use the same set of Wikipedia articles to construct the clustering ground truth where each section in the article corresponds to a subtopic, and passages within the same section are supposed to be clustered together in the adjacency matrix **T**.

The Wikipedia articles for the benchmark are collected from the TREC CAR passage retrieval dataset [3]. Specifically, we use *benchmarkY1train* and *benchmarkY1test* top-level datasets for evaluation and the separate large *train* dataset for training. Note that, articles in any of the benchmark datasets are removed from the large *train* dataset to avoid training data leakage. [2]

5.2 Experimental Setup

We explore the task of passage re-ranking to evaluate whether our clustering supervision helps the retrieval model achieve better results in terms of precision. Given the queries q and the sets of candidate passages P_q, the ORCA system produces a re-ranking of the passages.

Evaluation metric. Re-ranking results under each system is evaluated against our two test benchmarks in terms of mean average precision (MAP), normalized cumulative discounted gain (nDCG) and R-precision (Rprec).

[2]Code and relevant instructions to reproduce the results discussed in this paper are available here: https://github.com/nihilistsumo/ORCA.

Table 1: Comparison of lead passage retrieval performance in terms of mean average precision (MAP). The datasets used for this experiment are the benchmarkY1train and benchmarkY1test obtained from the TREC CAR dataset. Statistically significant improvement over monoBERT according to a standard error bar overlap test is denoted with $^{\triangle}$.

Methods	benchmarkY1train			benchmarkY1test		
	MAP	nDCG	Rprec	MAP	nDCG	Rprec
monoBERT	0.422	0.610	0.301	0.407	0.604	0.313
IDCM	0.467	0.648	0.351	0.452	0.651	0.372
ORCA-bin	0.452	0.632	0.331	0.427	0.625	0.324
ORCA	**0.483**$^{\triangle}$	**0.658**	**0.369**	**0.479**$^{\triangle}$	**0.671**	**0.383**

Compared systems. For the experimental results, we evaluate the following four models for the re-ranking task:

- **monoBERT:** This is the original monoBERT re-ranker [15] without any clustering supervision, but trained on the same training data.
- **IDCM:** This is representative of more recent and complex document re-rankers used for ad-hoc ranking tasks. Specifically, this model employs a cascading approach to filter relevant documents [5] in two stages with the later stage being more accurate but computationally expensive than the previous.
- **ORCA with subtopic clustering supervision:** This is our proposed approach; subtopic clustering ground truth is provided to the re-ranker as the clustering supervision.
- **ORCA-bin with binary clustering supervision:** Binary clustering ground truth is used as the clustering supervision. We include this method as a baseline to confirm RQ2.

5.3 Overall Re-Ranking Quality

Table 1 presents the evaluation results on the re-ranking task conducted on the two test benchmarks. We observe that our proposed ORCA approach with subtopic clustering supervision outperforms the unmodified monoBERT approach by a large margin. This observation is consistent for both *benchmarkY1train* and *benchmarkY1test*. This supports our hypothesis that providing supervision from a subtopic clustering model helps the retrieval system in identifying overview passages and hence answers the **RQ1** with "yes".

To answer **RQ2**, we explore the performance of ORCA-bin approach that receives binary clustering supervision. Although ORCA-bin improves upon the unmodified monoBERT, it is outperformed by our ORCA system with subtopic clustering supervision. This suggests that providing clustering supervision with subtopics is much more beneficial than binary clustering while training an embedding model for re-ranking tasks.

5.4 Evaluation on a Per-Query Basis

We perform a hurts-helps analysis, counting on how many queries our model obtains an improvement over the monoBERT baseline in terms of MAP. We observe that among 115 queries in

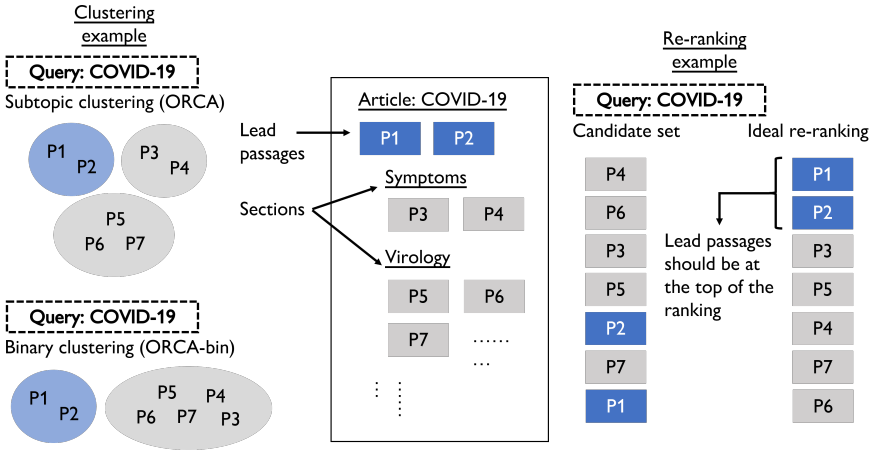


Figure 3: Center: Example TREC CAR benchmark article structure. All passages in the article P1 to P7 form the candidate set. The article title is considered as the query.
Right: The task is for a given query e.g. *COVID-19*, re-rank all passages in the candidate set such that the lead passages for the query (in this case P1 , P2) are on top of the ranking.
Left top: We provide clustering supervision to the embedding model so that the resulting embedding vectors would form subtopic clusters as shown.
Left bottom: An alternative way is to consider only two clusters with respect to the query: relevant and non-relevant. This is referred to as the binary clustering and used as a baseline in our evaluation.

the TREC CAR benchmarkY1train dataset, 73 queries (63%) obtain an improvement from the clustering supervision. For the benchmarkY1test, the improvement is even stronger, 92 out of 126 queries achieve a better MAP score than the unmodified monoBERT baseline. This suggests that supervision from a query-specific clustering model leads to better re-ranking performance for ranking overview passages.

5.5 Qualitative Analysis

The empirical evaluation shows that clustering supervision helps retrieval models in identifying overview passages. We further conduct a qualitative analysis on some example queries with their respective relevant passages to understand the salient qualities of the overview passages and why clustering supervision helps to identify them.

Query: Killifish

Relevant passages:

> The word killifish is of uncertain origin, but is likely to have come from the Dutch kil for a kill (small stream). Although killifish is sometimes used as an English equivalent to the taxonomical term Cyprinodontidae, some species belonging to that family have their own common names, such as the pupfish and the mummichog.

> A killifish is any of various oviparous (egg-laying) cyprinodontiform fish (including families Aplocheilidae , Cyprinodontidae , Fundulidae , Profundulidae and Valenciidae). Altogether, there are some 1270 different species of killifish, the biggest family being Rivulidae, containing more than 320 species. ...

For the query "Killifish", the above two passages are relevant according to the ground truth among 13 passages in the candidate set which are to be re-ranked. From the per-query analysis, we observe that for this query the ORCA system obtains a MAP of 0.94 while monoBERT only obtains a MAP of 0.4. As we observe for most of the queries in the benchmark, we can see that the corresponding relevant passages are about a much broader topic encompassing many different subtopics about the "Killifish" query ranging from etymology to its biological classifications. This is correctly identified by our ORCA system while monoBERT fails to do so.

6 CONCLUSION

Broad queries retrieve documents that span multiple subtopics around the query topic. Research works similar to TREC CAR aim to answer such queries through automatically generated articles from relevant Wikipedia passages. While passage retrieval for queries is a well-explored area in IR research, we focus on identifying overview passages from the candidate set. We propose the ORCA (Overview Retriever with Clustering Augmentation) system to identify such overviews given a query. This can be used to design search result pages that directly inform the user without any further navigation (e.g., good abandoment).

Inspired by the clustering hypothesis, we present Topic-Mono-Bert that uses supervision from a subtopic clustering model to help a retrieval model. Empirical evaluation on two Wikipedia-based benchmarks shows that the proposed model significantly outperforms monoBERT and IDCM, two recent neural retrieval systems, on the task of overview passage retrieval.

While we use monoBERT as an example of a strong neural ranker in this work, additional topic objective can be directly incorporated into any neural ranking system.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1846017. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Nir Ailon, Moses Charikar, and Alantha Newman. 2008. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)* 55, 5 (2008), 1–27.

[2] Laura Dietz, Shubham Chatterjee, Connor Lennox, Sumanta Kashyapi, Pooja Oza, and Ben Gamari. 2022. Wikimarks: Harvesting Relevance Benchmarks from Wikipedia. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3003–3012.

[3] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC Complex Answer Retrieval Overview.. In *TREC*.

[4] Jiyin He, Edgar Meij, and Maarten de Rijke. 2011. Result diversification based on query-specific cluster ranking. *Journal of the American Society for Information Science and Technology* 62, 3 (2011), 550–571.

[5] Sebastian Hofstätter, Bhaskar Mitra, Hamed Zamani, Nick Craswell, and Allan Hanbury. 2021. Intra-document cascading: learning to select passages for neural document ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1349–1358.

[6] Nick Jardine and Cornelis Joost van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information storage and retrieval* 7, 5 (1971), 217–240.

[7] Sen Jia, Guihua Tang, Jiasong Zhu, and Qingquan Li. 2015. A novel ranking-based clustering approach for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing* 54, 1 (2015), 88–102.

[8] Sumanta Kashyapi and Laura Dietz. 2022. Query-Specific Subtopic Clustering. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (Cologne, Germany) *(JCDL '22)*. Association for Computing Machinery, New York, NY, USA, Article 11, 9 pages. https://doi.org/10.1145/3529372.3530923

[9] Oren Kurland. 2009. Re-ranking search results using language models of query-specific clusters. *Information Retrieval* 12, 4 (2009), 437–460.

[10] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good abandonment in mobile and PC internet search. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 43–50.

[11] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies* 14, 4 (2021), 1–325.

[12] Robert Litschko, Federico Nanni, and Goran Glavas. 2018. Trec-CAR 2018: A Simple Unsupervised Semantic Query Expansion Model.. In *TREC*.

[13] Xiaoyong Liu and W Bruce Croft. 2004. Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 186–193.

[14] Sean MacAvaney, Nazli Goharian, Ophir Frieder, and Andrew Yates. 2018. PACRR Gated Expansion for TREC CAR 2018.. In *TREC*.

[15] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424* (2019).

[16] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375* (2019).

[17] Anastasios Tombros, Robert Villa, and Cornelis J Van Rijsbergen. 2002. The effectiveness of query-specific hierarchic clustering in information retrieval. *Information processing & management* 38, 4 (2002), 559–582.

[18] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Detecting good abandonment in mobile search. In *Proceedings of the 25th International Conference on World Wide Web*. 495–505.

[19] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Is this your final answer? evaluating the effect of answers on good abandonment in mobile search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 889–892.

[20] Xing Yi and James Allan. 2009. A comparative study of utilizing topic models for information retrieval. In *European conference on information retrieval*. Springer, 29–41.