2024

Evaluation of Hypothetical Document and Query Embeddings for Information Retrieval Enhancements in the Context of Diverse User Queries

Marten Jostmann
viadee Unternehmensberatung AG, Germany, marten.jostmann@viadee.de

Hendrik Winkelmann
viadee Unternehmensberatung AG, Germany, hendrik.winkelmann@viadee.de

Evaluation of Hypothetical Document and Query Embeddings for Information Retrieval Enhancements in the Context of Diverse User Queries

Research in Progress

Marten Jostmann and Hendrik Winkelmann

viadee Unternehmensberatung AG, Münster, Germany
{marten.jost,hendrik.winkel}mann@viadee.de

Abstract. The task of Information Retrieval (IR) is the foundation of question-answering and information search systems. The complexity of these retrieval systems evolved to enable an effective search of a variety of documents based on users' information needs. However, the quality of the retrieval results highly depends on the question phrasing and structure heavily influenced by the user's age, expertise, background, and aim. This paper aims to close the gap between user-specific query structures and the content of various documents by employing the approaches Hypothetical Document Embeddings (HyDE) and Hypothetical Query Embeddings (HyQE). Both approaches achieve promising results in terms of effectiveness and robustness. An evaluation indicates that HyDE and HyQE outperform traditional retrieval methods such as BM25 and plain embedding methods.

Keywords: Information Retrieval, Diverse Queries, HyDE, HyQE

1 Introduction

Large Language Models (LLMs) have demonstrated an extraordinary linguistic understanding and the capability of human-like text generation, reasoning, and question-answering (Gatt & Krahmer (2018), Wei et al. (2022), Jiang et al. (2021)). In addition, LLMs are able to gain and reproduce knowledge through extensive pre-training (Petroni et al. (2019)). However, access to new knowledge is limited and prone to hallucinations, i.e., nonfactual and wrong information (Lewis et al. (2020), Bang et al. (2023)). Hence, LLMs should not be utilized directly in organizational settings or critical domains. Recently, Retrieval-Augmented Generation (RAG) has gained attention as a possibility to solve this problem. However, the basis of all RAG systems is an underlying Information Retrieval (IR) the quality of which depends strongly on the question phrasing and structure heavily influenced by the age, expertise, background, and aim of the user (Bilal & Kirby (2002), Penha et al. (2022)), thus requiring further tailoring. This research aims to close the gap between query structures and phrasing and the content of different documents in large collections, hence enabling effective and robust RAG. Hypothetical Document Embeddings (HyDE) and Hypothetical Query Embeddings (HyQE) are investigated in the presence of query variations. Both approaches aim to bridge the lexical gap between the different phrasings of queries and documents by leveraging LLMs.

19th International Conference on Wirtschaftsinformatik
September 2024, Würzburg, Germany

2 Approach

Hypothetical Document Embeddings (HyDE) and Hypothetical Query Embeddings (HyQE) are alternatives for approaching the concerns of information phrasing and structure. While HyDE is already described in previous work, HyQE is a combination of old ideas and new technology (see Nogueira, Yang, Lin & Cho (2019), Nogueira, Lin & Epistemic (2019), Ma et al. (2021), Bonifacio et al. (2022)).

Hypothetical Document Embeddings Empirically, HyDE has already shown to improve IR results over traditional approaches (Gao et al. (2023)). HyDE can be used in various domains without requiring a labeled dataset. The core idea is the generation of a hypothetical answer document given a query. The generated document is expected to represent the target documents more effectively than the query itself. Thus, using it for search is expected to increase the effectiveness of IR.

An important aspect not considered in the original publication of HyDE is the handling of query variations and diverse user groups. Still, the creation of hypothetical documents is promising to mitigate these challenges. Compared to methods such as BM25 and dense embeddings, HyDE overcomes the lexical chasm between query and document. This is done by using the language comprehension and reasoning capabilities of LLMs to capture the different linguistic nuances of a query and expand its context.

Furthermore, the vocabulary probability distributions of queries and documents can be aligned by generating hypothetical documents. For instance, question-specific terms such as *what*, *when*, *where*, *why*, and *who* occur less frequently in documents but rather often in questions (Berger et al. (2000)). Such words would be deemphasized by the context expansion of LLMs, while the content to be searched for is being prioritized.

Hypothetical Query Embeddings A major disadvantage of HyDE is the unawareness of the document corpus. Using the HyDE approach it can never be assumed with absolute certainty whether the generated document is relevant to the given query or not (see Lewis et al. (2020)). This problem can be avoided by using HyQE since the process is reversed. Instead of generating hypothetical documents that imitate relevant documents, HyQE tries to generate suitable queries for a specific document. Queries are generated during the indexing phase, enabling an efficient IR through the use of dense embeddings. Compared to the HyDE approach, the user query only has to be embedded by an embedding model, leading to a time- and cost-efficient retrieval phase. In Figure 1 the indexing process of HyQE is depicted.

Initially, the potential target documents have to be split into smaller pieces using an appropriate chunking technique. For each text chunk, an LLM model can generate multiple hypothetical queries. Two options are conceivable. (1) Split generated queries and embed them separately and (2) concatenate all generated queries with the respective chunk. This outcome then is embedded. In the first variant, each query is weighted higher and, therefore, has a greater influence on the IR results in the case of a matching user query. The comparison of hypothetical queries and actual user queries is much more natural and associated with less bias. However, the missing document context can decrease IR performance. This can be mitigated using the second approach where

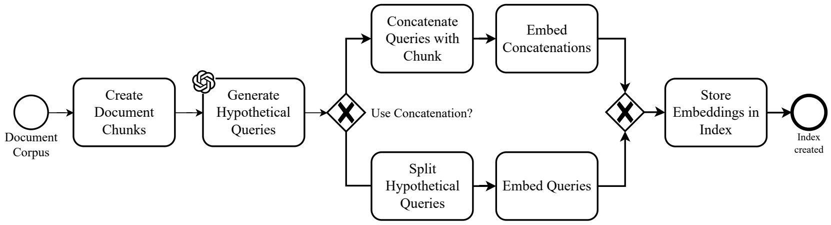

Document Corpus
Create Document Chunks
Generate Hypothetical Queries
Use Concatenation?
Concatenate Queries with Chunk
Embed Concatenations
Split Hypothetical Queries
Embed Queries
Store Embeddings in Index
Index created

Figure 1. HyQE indexing process

all queries are concatenated with the respective chunk from the target document, thus preserving the document's context.

Moreover, HyQE can also prove advantageous in managing query variations in syntax and semantics: HyQE allows for the generation of specific hypothetical queries in various linguistic variants for the same information need. The underlying information need does not have to be inferred from ambiguous queries but can rather be directly used as the basis for creating questions. In addition, it is possible to create queries from different user perspectives and with consideration of the user's prior knowledge. The effectiveness of the subsequent IR system therefore depends on the number of generated queries, as the broadest coverage of possible variations can only be achieved by scaling up the query amount.

As all time-consuming and expensive steps are processed in the upstream indexing process, the downstream retrieval task can be processed rather quickly. The user is entering a query into the IR system which can be directly embedded without further preprocessing. The next step is to load the respective index from the indexing database. Subsequently, the user embedding is compared to each embedding in the index using a similarity or distance function.

3 Experiments

3.1 Setup

Implementation Both HyDE and HyQE do not depend on a specific LLM or embedding model. For the experiments the `gpt-3.5-turbo-1106` model from OpenAI (see Brown et al. (2020)) with a temperature of 0.3 and a frequency penalty of 0.5 was used to generate hypothetical documents and queries. This generation process was limited to a maximum of 1 500 tokens, allowing the model to freely generate documents and queries but still maintaining a reasonable time and cost effort. The `text-embedding-ada-002` was used as the embedding model (see Neelakantan et al. (2022)). The term-based method BM25 and semantic search with plain embeddings serve as the baseline to compare it to the new approaches.

Datasets Both approaches were evaluated using two real-world test collections out of different domains. The first dataset, abbreviated as DS_1, is an excerpt of the internal

Confluence document database of the viadee IT consultancy with over 9400 documents that are different from each other in terms of their subject. The second dataset (DS_2) consists of 30 HTML and PDF files from a public law institute in Germany. Both datasets consist of documents in German and are queried by diverse user groups with different information needs and each represents a different domain.

In order to perform a concise evaluation, not only datasets are required but also test collections containing different queries with their ground truth target documents. A relevance label is defined for each document based on the four-point scale of the TREC 2021 test collection (Craswell et al. (2022)). The first test collection for the Confluence dataset consists of eleven topics with a total of 83 queries. The second test collection contains ten topics with 18 queries.

For the robustness evaluation, both collections were extended with automatically generated query variations based on the variations observed by PENHA ET AL. (2022).

Metrics The effectiveness of the retrieval approaches will be measured using the normalized Discounted Cumulative Gain (nDCG) (see Järvelin & Kekäläinen (2002)). Furthermore, a measure of the retrieval robustness is required. The robustness evaluation should focus on the degree to which the IR system produces consistent outcomes in the presence of query variations. Hence, by considering only the robustness of the system the retrieval quality is neglected. One possibility to measure the robustness is the Inter-Topic Variance (ITV). For this purpose, the variance of the result of any effectiveness metric Φ_i^n is calculated for each topic $n \in \{1, ..., T\}$ consisting of several queries $i \in \{1, ..., Q\}$. Lastly, the mean Inter-Topic Variance (mITV) is calculated as shown in Equation 1.

$$mITV = \frac{1}{T}\sum_{n=1}^{T}\sigma_n^2, \quad \text{where } \sigma_n^2 = \frac{1}{Q}\sum_{i=1}^{Q}(\Phi_{n,i} - \bar{\Phi}_n)^2 \tag{1}$$

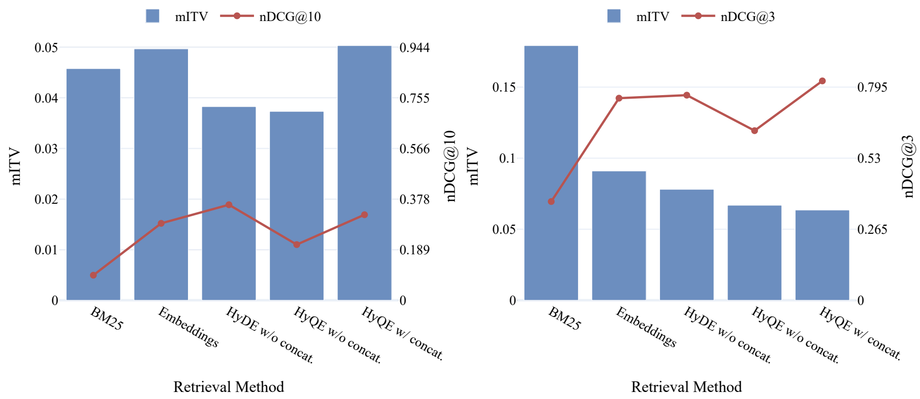
In the later experiments, the nDCG score is used as the metric Φ, since it is the most informative measurement. A low mITV suggests a more robust system that can handle diverse user queries, as it indicates a reduced variability within the same information need. However, in order to be able to evaluate the overall performance of the IR system, the effectiveness must always be considered.

3.2 Results

In Figure 2 the effectiveness and robustness of various approaches including the `BM25` and plain `Embeddings` baseline is depicted for two real-world datasets. For DS_1 (see Figure 2a) both metrics are calculated based on the top 10 retrieval results. Due to the reduced size of DS_2 (see Figure 2b), only the top 3 retrieval results will be considered.

The results in Figure 2 indicate that `BM25` is a solid baseline but performs poorly compared to the `Embeddings` baseline. The overall best effectiveness could be achieved by the HyDE approach for the DS_1 test collection. However, for the DS_2 test collection, HyQE with concatenation was able to achieve an `ndcg@3` score of over 80 %.

Concerning the robustness, the results indicate that both baseline approaches are highly sensitive to query variations. By using HyDE, the robustness could be signif-



(a) DS_1 test collection

(b) DS_2 test collection

Figure 2. Effectiveness and robustness of IR approaches

icantly increased compared to the baseline. Simultaneously, the effectiveness of the retrieval results is improved. It is notable that both HyQE variants exhibit significantly stronger deviations across different test collections. For example, the HyQE concatenation approach has both the highest robustness and the highest effectiveness on the second test collection, whereas the same approach has the lowest robustness on the first test collection. HyQE without concatenation exhibits the highest robustness on the DS_2 with the drawback of a lower effectiveness. Irrespective of the robustness deviations, HyQE with concatenation proves to be significantly more effective as the alternative approach in all experiments using Ada embeddings.

3.3 Discussion

The experiments on the effectiveness indicated a strong performance improvement when using HyDE and HyQE with Ada embeddings. The strengths of the HyDE approach originally proposed by Gao et al. could be validated. In addition to the general improvement in the document ranking, HyDE was able to find relevant documents for several queries and topics. Moreover, the HyQE approach demonstrated that the generation of potential user queries can result in a better ranking of relevant documents. The results indicate that the generated queries have to be concatenated with the document to achieve better rankings since the similarity search then can take the document context into account. As the two datasets stem from different domains, generalizability was validated to a certain degree. We expect HyQE to excel in domains where a diverse audience should be able to search for information containing domain-specific language using plain language, as is the case, e.g., in the medical and legal domains. On the other hand, HyDE might benefit from standard document structures in these domains, retrieving required sections.

In addition to the effectiveness, the robustness of the different methods also played a major role in the evaluation. The experiments showed that HyDE was able to deliver constant results even in the presence of various query variations. Contrary to the expectations, HyQE was not able to always improve the robustness of the IR system. While the robustness of HyQE was very low on the DS_1 test collection, it showed the highest robustness on the DS_2 test collection. Overall, the robustness of HyQE strongly depends on the selected dataset, the generated queries as well as the embedding method. Hence, no overarching conclusion can be inferred regarding the effectiveness and robustness of individual retrieval methods. It is promising to further investigate the effect of prompt engineering as a domain-specific hyperparameter tuning-step.

4 Related Work

Traditional retrieval methods based on near-exact match search, such as TF-IDF and BM25, often suffer from the lexical gap (Berger et al. (2000)). To address this limitation, dense representations in latent semantic spaces have been proposed (Zhao et al. (2024)). In contrast to a sparse representation, dense text embeddings map the text into a fixed-sized, rather low-dimensional vector space. Given a query, similar or relevant texts can be found by comparing the similarity of these embedding vectors.

Based on dense text embeddings, query and document expansion mechanisms aim to refine user queries and documents to enhance the retrieval effectiveness (Billerbeck & Zobel (2005)). With the help of query expansion techniques, the query can be extended with additional query terms to minimize the lexical gap between the query and the document. Popular approaches are Doc2Query and Inpars (Nogueira, Yang, Lin & Cho (2019), Bonifacio et al. (2022)). In contrast to document expansion, query expansion shifts the focus from extending documents to refining queries. The most prominent example is *RM3* which iteratively adapts the user query by considering the retrieved documents as an additional query context (Abdul-Jaleel et al. (2004)). A promising approach leveraging the generative capabilities of LLMs was proposed by GAO ET AL. (2023), namely HyDE. Instead of extending the query with related terms, they suggest generating an intermediate hypothetical document that serves as a better comparison basis.

5 Conclusion

The main goal of this paper was to develop and apply HyDE and HyQE for IR in the context of diverse user queries. In particular, the effectiveness and robustness of these approaches were compared to traditional retrieval methods. The experimental results showed a high effectiveness and robustness of HyDE using both datasets compared to the baseline. However, hypothetical document generation is costly and time-consuming, which is the major drawback of this approach. HyQE mitigates this shortcoming by enabling a faster as well as cheaper retrieval but comes with the drawback of a more expensive indexing phase. The evaluation revealed a high effectiveness for the HyQE concatenation variant. The LLM model was able to generate realistic queries, considering different user perspectives and query types, which led to an increase in the ranking of relevant documents.

References

Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., Smucker, M. D. & Wade, C. (2004), 'Umass at trec 2004: Novelty and hard', *Computer Science Department Faculty Publication Series* p. 189.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y. & Fung, P. (2023), A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, *in* J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti & A. A. Krisnadhi, eds, 'Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Nusa Dua, Bali, pp. 675–718.
URL: *https://aclanthology.org/2023.ijcnlp-main.45*

Berger, A., Caruana, R., Cohn, D., Freitag, D. & Mittal, V. (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', pp. 192–199.

Bilal, D. & Kirby, J. (2002), 'Differences and similarities in information seeking: children and adults as web users', *Information Processing & Management* **38**(5), 649–670.
URL: *https://www.sciencedirect.com/science/article/pii/S0306457301000577*

Billerbeck, B. & Zobel, J. (2005), 'Document expansion versus query expansion for ad-hoc retrieval', *ADCS 2005 - Proceedings of the Tenth Australasian Document Computing Symposium* .

Bonifacio, L., Abonizio, H., Fadaee, M. & Nogueira, R. (2022), 'Inpars: Data augmentation for information retrieval using large language models'.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020), 'Language models are few-shot learners', *Advances in neural information processing systems* **33**, 1877–1901.

Craswell, N., Mitra, B., Yilmaz, E., Campos, D. & Lin, J. (2022), Overview of the trec 2021 deep learning track, *in* 'Text REtrieval Conference (TREC)', TREC.
URL: *https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2021-deep-learning-track/*

Gao, L., Ma, X., Lin, J. & Callan, J. (2023), Precise zero-shot dense retrieval without relevance labels, *in* A. Rogers, J. Boyd-Graber & N. Okazaki, eds, 'Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1762–1777.

Gatt, A. & Krahmer, E. (2018), 'Survey of the state of the art in natural language generation: Core tasks, applications and evaluation', *Journal of Artificial Intelligence Research* **61**, 65–170.

Jiang, Z., Araki, J., Ding, H. & Neubig, G. (2021), 'How can we know when language models know? on the calibration of language models for question answering', *Transactions of the Association for Computational Linguistics* **9**, 962–977.

Järvelin, K. & Kekäläinen, J. (2002), 'Cumulated gain-based evaluation of ir techniques', *ACM Trans. Inf. Syst.* **20**(4), 422–446.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S. & Kiela, D. (2020), Retrieval-augmented generation for knowledge-intensive nlp tasks, *in* 'Proceedings of the 34th International Conference on Neural Information Processing Systems', NIPS'20, Curran Associates Inc, Red Hook, NY, USA.

Ma, J., Korotkov, I., Yang, Y., Hall, K. & McDonald, R. (2021), Zero-shot neural passage retrieval via domain-targeted synthetic question generation, *in* P. Merlo, J. Tiedemann & R. Tsarfaty, eds, 'Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume', Association for Computational Linguistics, Online, pp. 1075–1088.
URL: *https://aclanthology.org/2021.eacl-main.92*

Neelakantan, A., Weng, L., Power, B. & Jang, J. (2022), 'Introducing text and code embeddings'.
URL: *https://openai.com/blog/introducing-text-and-code-embeddings*

Nogueira, R., Lin, J. & Epistemic, A. I. (2019), 'From doc2query to docTTTTTquery', *Online preprint* **6**, 2.

Nogueira, R., Yang, W., Lin, J. & Cho, K. (2019), 'Document expansion by query prediction'.

Penha, G., Câmara, A. & Hauff, C. (2022), Evaluating the robustness of retrieval pipelines with query variation generators, *in* M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg & V. Setty, eds, 'Advances in Information Retrieval', Vol. 13185 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, pp. 397–412.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y. & Miller, A. (2019), Language models as knowledge bases?, *in* K. Inui, J. Jiang, V. Ng & X. Wan, eds, 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 2463–2473.
URL: *https://aclanthology.org/D19-1250*

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le Quoc & Zhou, D. (2022), 'Chain-of-thought prompting elicits reasoning in large language models'.
URL: *http://arxiv.org/pdf/2201.11903v6*

Zhao, W. X., Liu, J., Ren, R. & Wen, J.-R. (2024), 'Dense text retrieval based on pretrained language models: A survey', *ACM Trans. Inf. Syst.* **42**(4).