Query Generation with External Knowledge for Dense Retrieval

Sukmin Cho Soyeong Jeong Wonsuk Yang Jong C. Park[*]
School of Computing
Korea Advanced Institute of Science and Technology
{nelllpic,syjeong,derrick0511,park}@nlp.kaist.ac.kr

Abstract

Dense retrieval aims at searching for the most relevant documents to the given query by encoding texts in the embedding space, requiring a large amount of query-document pairs to train. Since manually constructing such training data is challenging, recent work has proposed to generate synthetic queries from documents and use them to train a dense retriever. However, compared to the human labeled queries, synthetic queries do not generally ask for hidden information, therefore leading to a degraded retrieval performance. In this work, we propose *Query Generation with External Knowledge* (QGEK), a novel method for generating queries with external knowledge related to the corresponding document. Specifically, we convert a query into a triplet-based template to accommodate external knowledge and transmit it to a pre-trained language model (PLM). We validate QGEK in both in-domain and out-domain dense retrieval settings. The dense retriever with the queries requiring external knowledge is found to make good performance improvement. Also, such queries are similar to the human labeled queries, confirmed by both human evaluation and unique & non-unique words distribution.

1 Introduction

Information retrieval (IR) is the task of collecting relevant documents from a large corpus when given a query. IR not only plays an important role in the search system by itself, but is also crucially applied to various NLP tasks such as Open-Domain QA (Kwiatkowski et al., 2019) and Citation-Prediction (Cohan et al., 2020) with its ability to find grounding documents. As the simplest retrieval method, traditional term-based sparse models such as TF-IDF and BM25 (Robertson and Zaragoza, 2009) are widely used. However, these sparse retrieval models are unable to capture the semantic similarities without explicit lexical overlaps between the query and its relevant documents. As a solution, dense retrieval models are recently proposed where query and document representations are embedded into the latent space (Gillick et al., 2018; Karpukhin et al., 2020), though they require a large amount of paired query-document training samples for notable performance, which is very challenging and expensive. In response, a zero-shot setting is often adopted, but dense retrievers are known to show poor performance on a new target domain (Ma et al., 2021; Wang et al., 2021; Xin et al., 2021).

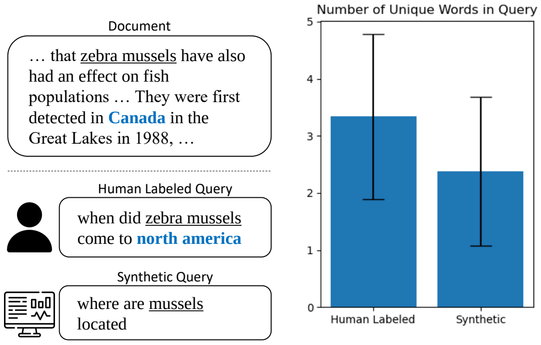

Figure 1: The analysis of human labeled query and synthetic query. (Left) Examples of the human labeled query and synthetic query. (Right) Average of unique words in human labeled query and synthetic query.

One possible solution is to generate synthetic queries by fine-tuning a pre-trained language model (PLM) on a large IR benchmark dataset, and to use such queries for training dense retrievers (Ma et al., 2021; Thakur et al., 2021; Wang et al., 2021). However, this method does not yet provide synthetic queries whose quality is comparable to that of human labeled ones, thus hindering retrieval performance.

In particular, we argue that, for the effective training of dense retrievers, query samples should be allowed to contain external knowledge that is not explicitly shown in documents. As Figure 1

[*] Corresponding author

Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 22 - 32
May 27, 2022 ©2022 Association for Computational Linguistics

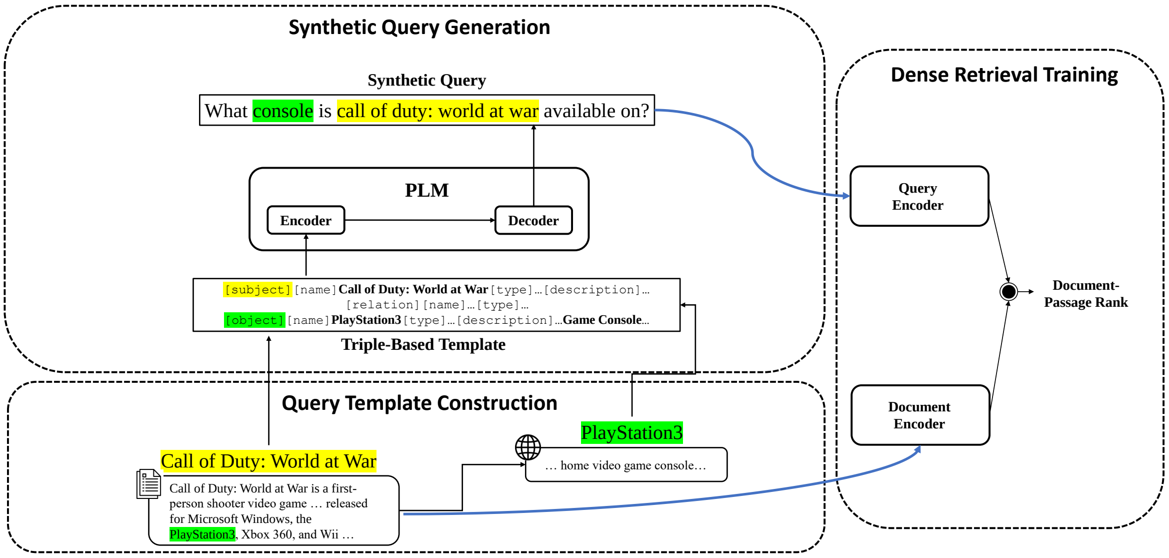
Synthetic Query Generation
Synthetic Query
What console is call of duty: world at war available on?
PLM
Encoder
Decoder
[subject] [name] Call of Duty: World at War [type]…[description]…
[relation] [name]…[type]…
[object] [name] PlayStation3 [type]…[description]…Game Console…
Triple-Based Template
Query Template Construction
PlayStation3
… home video game console…
Call of Duty: World at War
Call of Duty: World at War is a first-person shooter video game … released for Microsoft Windows, the PlayStation3, Xbox 360, and Wii …
Dense Retrieval Training
Query Encoder
Document Encoder
Document-Passage Rank

Figure 2: Overall methods of query generation with external knowledge and dense retrieval training with synthetic queries.

shows, the human labeled query contains the external knowledge that Canada and North America are related, which is easily grasped by humans but not by the machine. Also, unique words in the query, often considered as external knowledge, are more frequently included in the human labeled queries than in the synthetic queries. The dense retrievers would better capture semantic relations if they are trained with such queries that show more characteristics of human labeled ones.

In this paper, we focus on generating queries with external knowledge by employing a simple method of explicitly transmitting document-related information to a PLM. Even though PLMs can handle hidden information to some extent by learning from a large amount of data, we argue that transmitting additional pieces of external knowledge to a PLM contributes positively to generating queries requiring external knowledge. Specifically, we first interpret the given query into a triplet-based template to consider the given document and related external knowledge together. A PLM is then fine-tuned to generate queries from triplet-based templates, together with a processed KB-QA dataset. The dense retriever is trained with the synthetic queries from the template extracted from the given document and corresponding external knowledge. The proposed method, henceforth referred to as *Query Generation with External Knowledge* (QGEK), is schematically illustrated in Figure 2.

We validate QGEK in both in-domain and out-domain (zero-shot) dense retrieval settings with diverse evaluation metrics. The experimental results show that queries that require external knowledge to answer are helpful for improving retrieval performance. Furthermore, we provide detailed qualitative analyses of synthetic queries and discuss which aspects of queries should be considered when training dense retrieval models.

Our contributions in this work are threefold:

- We propose a generation method of queries that require hidden information, not present in the document, from external sources.
- We experimentally show that the generated queries are similar to the gold queries that are labeled by human annotators.
- We evaluate the quality of generated queries with respect to dense retrieval performance and distribution of unique words so as to find optimal queries in training a dense retriever.

2 Related Work

2.1 Dense Retriever

The sparse retriever, a traditional IR system, retrieves the target documents based on the lexical values such as frequency of terms and documents. BM25 (Robertson and Zaragoza, 2009) has been arguably the most frequently used method for such IR. However, as the retriever mainly handles the match of the lexical entries, ‘semantically similar’ but not the same lexical entries are not considered in the search for documents, affecting the user experience (Berger et al., 2000).

The dense retriever (Karpukhin et al., 2020) has received much attention as a solution to handle the problem, triggered by the Transformer (Vaswani et al., 2017) network and PLM. A dense retriever fetches the documents located closest to the query vector in the dense vector space with the results recorded in advance for retrieval performance. The model maps queries and documents to the dense vector space using a bi-encoder structure initialized from a PLM such as BERT (Devlin et al., 2019a).

The dense retriever requires a large-scale dataset for model training, and curating such datasets is a much arduous endeavor. Thakur et al. (2021) proposed a zero-shot setting where dense retrievers are trained on a single large IR corpus, rather than on every dataset. Nonetheless, retrieval in such setting is still quite challenging.

2.2 Query Generation

Query generation is a simple method that addresses the shortage of training data for a dense retriever (Ma et al., 2021; Thakur et al., 2021; Wang et al., 2021). The most commonly used method has been to fine-tune the T5-base model (Raffel et al., 2020) to the MS MARCO dataset (Nguyen et al., 2016) and create a synthetic query in the target domain. Exploiting the size and domain of MS MARCO, we can obtain an effective retrieval performance by fine-tuning the T5 model. Info-HCVAE (Lee et al., 2020) achieved good performance by designing the relationship between document, query, and answer as a probability distribution and learning the latent vectors based on an auto-encoder. Answers and documents are used as inputs when creating queries. In these two methods, however, the processing of hidden information in the document still depends only on PLMs.

The existing methods focus only on the given document when generating queries, without much consideration of hidden information. In contrast, QGEK includes not only the document but also the hidden information that can be inferred from the given document with external knowledge.

2.3 Exploiting External Knowledge

External knowledge has been widely used along with PLMs for several NLP tasks. (Wang et al., 2020) augmented PLMs using ConceptNet (Speer et al., 2017) for a commonsense question answering (QA) task and showed that KB, such as ConceptNet, contributes to the explicit grounding of the output, resulting in better reasoning abilities.

Furthermore, Zhou et al. (2018) proposed to generate knowledge-based dialogues for an Open-Domain Dialogue system. Dinan et al. (2019) confirmed that the additional external knowledge positively affects dialogue generation. In addition, Shuster et al. (2021) showed that the related external knowledge can be exploited to address critical issues, such as factual incorrectness and hallucination, in dialogue systems.

While external knowledge from KB has proved helpful in Commonsense QA and Open-Domain Dialogue domains, it is relatively underexplored for generating synthetic queries for dense retrieval. In this work, we adopt KB into a PLM for query generation and show the effectiveness of training dense retrievers with the synthetic queries on IR benchmark datasets.

3 Methods

QGEK is designed to generate a new synthetic query that requires an implicit inference process for the answer by exploiting both the given document and external knowledge hidden in the document. First, we interpret the query as the triplet <S,R,O> that can easily utilize both of them, where the triplet is converted into a single-text template to simplify the transmission to a PLM. Then, we construct triplet-based template & query pairs as training datasets for fine-tuning a PLM. For generating a query from target documents, the triplet-based template is extracted from a general document.

3.1 Preliminaries

The dense retriever maps query q and document d into an n-dimensional vector space with query encoder $E_Q(\cdot, \theta_q)$ and document encoder $E_D(\cdot, \theta_d)$ where θ is the encoder's parameter. The similarity score $f(q, d)$ between query q and document d is computed as a dot product:

$$f(q, p) = E_Q(q, \theta_q)^T \cdot E_D(d, \theta_d)$$

Training the dense retriever targets the vector space of which the relevant query and document pairs have a high similarity score compared to irrelevant pairs. Given query q, let (D_q^+, D_q^-) be the pairs of the sets of relevant documents and irrelevant documents. The objective function of dense

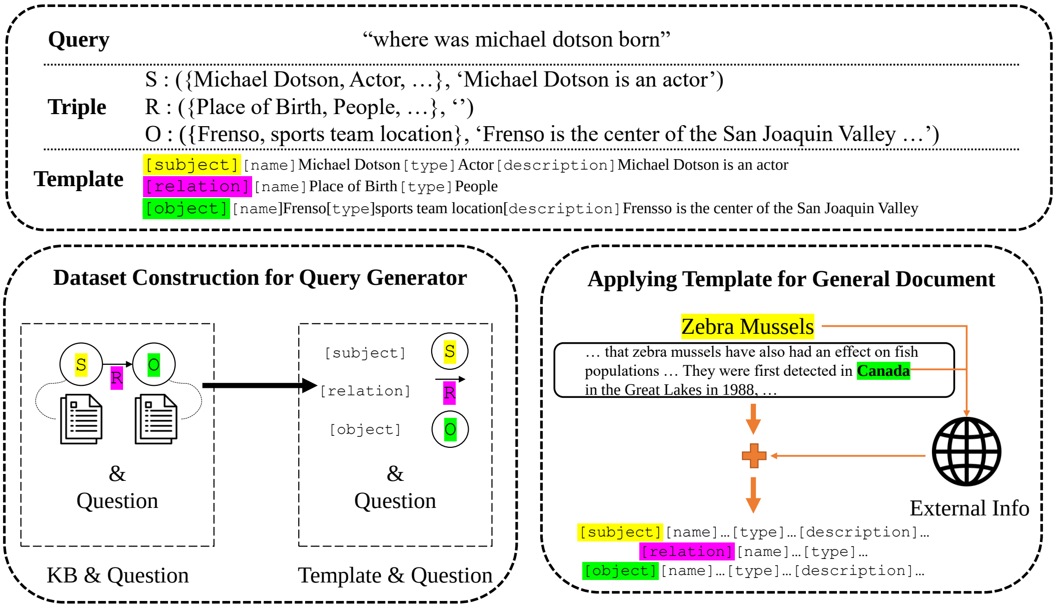
Query
“where was michael dotson born”
Triple
S : ({Michael Dotson, Actor, …}, ‘Michael Dotson is an actor’)
R : ({Place of Birth, People, …}, ‘’)
O : ({Frenso, sports team location}, ‘Frenso is the center of the San Joaquin Valley …’)
Template
[subject] [name] Michael Dotson [type] Actor [description] Michael Dotson is an actor
[relation] [name] Place of Birth [type] People
[object] [name]Frenso[type]sports team location[description] Frensso is the center of the San Joaquin Valley
Dataset Construction for Query Generator
S
R
O
& Question
[subject]
[relation]
[object]
& Question
KB & Question
Template & Question
Applying Template for General Document
Zebra Mussels
… that zebra mussels have also had an effect on fish populations … They were first detected in Canada in the Great Lakes in 1988, …
External Info
[subject] [name] … [type] … [description] …
[relation] [name] … [type] …
[object] [name] … [type] … [description] …

Figure 3: Overview of the Methods for Query Generation based on Triplet-based Template.

retriever is as follows:

$$\min_{\theta} \sum_{q} \sum_{d^{+}\in D_q^{+}} \sum_{d^{-}\in D_q^{-}} L(f(q,d^{+}), f(q,d^{-}))$$

The loss L is the negative log likelihood of the positive passage.

3.2 Query Interpretation as Triplet Form

Queries can be simply mapped to the $<S, R, O>$ triplet. The query $<S, R, O>$ asks for the answer O, which has relationship R with subject S. For example, the query "big little lies season 2 how many episodes" and the answer "7 episodes" can be mapped to <"big little lies season 2", "number of episodes", "7 episodes">.

Information of each item in a triplet can be largely divided into sentences or words units. We use two types of information to express each item of a triplet in more detail. Let $W_x = \{w_x^1, \dots, w_x^n\}$ and l_x be the set of word unit information and the single sentence unit information of the item x, respectively. Then, query Q can be interpreted as the triplet items with their own information:

$$Q = \{(W_S, l_S), (W_R, l_R), (W_O, l_O)\}$$

For generating a query that requires an implicit inference, a form of query that can utilize both the document and external knowledge is required. The proposed triplet simply handles both document and external knowledge by arranging information into the appropriate positions in the triplet. When transmitting such triplets to a PLM, we use the simple form of a single text template. The triplet-based template consists of triplet items delimited by special tokens as shown in Figure 3.

3.3 Dataset Construction for Query Generator

We construct a dataset consisting of triplet-query pairs for fine-tuning PLM. The KB based query can be converted into the proposed triplet. A canonical logical form of a KB based query is a representation that expresses the same meaning as the relationship between entities in KB. A simple interpretation of the proposed triplet can be seen as a canonical form consisting of two entities and a relationship between them.

For example, suppose that the entity, ‘Michael Dotson’, is first selected as subject S and has word unit information, ‘Actor’, and sentence unit information, ‘Michael Dotson is an actor’. Suppose also that there is an entity, ‘Frenso’, linked by ‘place-of-birth’ relationship with ‘Michael Dotson’. The other entity and relationship may have their own information from KB. The triplet-based template is created by combining all of them.

3.4 Applying Template for General Document

The fine-tuned PLM with the dataset constructed in Section 3.3 needs the triplet-based template to generate a query from a general document. We extract triplet items from the given document, and collect external knowledge to fill the template from the open web.

For example, suppose that there is a document about zebra mussels (cf. Figure 3). The subject S, relation R and object O are selected as 'zebra mussels', 'location' and 'Canada', respectively. The document alone is not enough to fill the information of object O, 'Canada'. The external knowledge, 'Canada is a country in North America', is extracted from the open web. Both given document and external knowledge are arranged into the appropriate positions in the template.

4 Experimental setups

We evaluate the performances of the dense retriever when trained with the synthetic queries compared to the human labeled queries. The dense retriever used in our experiments is the DPR (Karpukhin et al., 2020). The train dataset of the dense retriever is the pairs of the documents of Natural Question (NQ) (Kwiatkowski et al., 2019), also exploited as the source of the query generator, and the synthetic queries of the proposed method.

4.1 Datasets

We evaluate the effectiveness of the generated queries when using external knowledge on IR benchmark datasets. We conduct experiments in two settings: in-domain and out-domain (zero-shot). We measure the in-domain performance on the NQ and the out-domain performance on 13 representative IR datasets (Thakur et al., 2021).

In-Domain Dataset NQ (Kwiatkowski et al., 2019) is a benchmark dataset for the open-domain question answering task, fetched by Google search engine and from Wikipedia. We use the preprocessed version of the NQ following DPR (Karpukhin et al., 2020), which includes 58,880 training pairs and 7,405 test queries. The documents in NQ is used as input of query generator.

Out-Domain Dataset To validate the quality of generated queries for training the dense retriever, it is necessary to show the retrieval performance of diverse tasks. Each dataset used in out-domain experiments has diverse tasks and domains and requires retrieval models for finding grounding documents. They are shown in Table 1.

Task	Domain	Dataset
Argument Retrieval	Misc. Misc.	ArguAna (Wachsmuth et al., 2018) Touche-2020 (Bondarenko et al., 2020)
Entity-Retrieval	Wikipedia	DBPedia (Hasibi et al., 2017)
Question Anwering	Wikipedia Finance	HotpotQA (Yang et al., 2018) FiQA-2018 (Maia et al., 2018)
Duplicate-Question Retrieval	Quora	Quora (Thakur et al., 2021)
Fact Checking	Wikipedia Wikipedia Scientific	FEVER (Thorne et al., 2018) Climate-Fever (Leippold and Diggelmann, 2020) SciFact (Wadden et al., 2020)
Passage-Retrieval	Misc.	MS MARCO (Nguyen et al., 2016)
Citation-Prediction	Scientific	SCIDOCS (Cohan et al., 2020)
Bio-Medical IR	Bio-Medical Bio-Medical	TREC-COVID (Voorhees et al., 2021) NFCorpus (Boteva et al., 2016)

Table 1: Datasets for Out-Domain Experiments

4.2 Metrics

We explain the metrics for evaluating the performance of a dense retriever. In the basic setting, the retriever searches for top k relevant documents on a given query. We employ 4 metrics for top k documents: ACC@k, MRR@k, MAP@k, and nDCG@k. The in-domain experiment is evaluated with these 4 metrics, and the out-domain performance is evaluated with only nDCG@10.

ACC@k is the percentage of whether the correct documents are included in the top-k hits. It ignores the rank of retrieved documents.

MRR@k (Mean Reciprocal Rank) computes the average of the ranks of the first correct document from top-k documents. The rest of the correct documents are not included in computing MRR.

MAP@k (Mean Average Precision) first computes the average precision score of the correct documents' ranks in top-k hits for a given query. The mean of the average precision scores is the value of the MAP@K.

nDCG@k (Normalised Cumulative Discount Gain) is similar to MAP@k, but reflects the fact that the more relevant document is the more highly ranked in top-k documents.

4.3 Implementation Details

Query Generator We used BART (Lewis et al., 2020), one of the widely used PLMs, to generate the synthetic query from the proposed template. BART based on the transformer seq2seq architecture is trained by reconstructing text from noised input. The de-noising ability of BART is suitable for generating queries from text with noise from the external source.

SimpleQuestions (Bordes et al., 2015) (SQ), a question answering dataset based on KB, is se-

In-Domain

Train Query (↓)	**ACC@10**	**ACC@100**	**MRR@10**	**MRR@100**	**MAP@10**	**MAP@100**	**NDCG@10**	**NDCG@100**
Gold	.6374	.8974	.3372	.3493	.3146	.3296	.3892	.4543
QGEK	.4901	.7488	.2375	.2484	.2220	.2354	.2841	.3449
(-) Ext. Knowledge	.4860	.7285	.2348	.2457	.2162	.2295	.2745	.3357

Out-Domain

Train Query (↓)	**Arguana**	**DBpedia**	**fiqa**	**HotpotQA**	**NFC**	**Quora**	**SciFact**	**Touche-2020**	**C-Fever**	**Fever**	**MS MARCO**	**SciDocs**	**TREC-COVID**	**Avg.**
Gold	.2203	.2585	.1763	.3205	.2226	.5778	.4476	.2334	.1609	.5160	.1858	.1022	.5152	.3029
QGEK	.0948	**.2733**	.1182	**.4226**	.1791	.4982	.3436	.2093	**.1754**	**.7731**	.1738	.0945	.4411	.2921
(-) Ext. Knowledge	.0568	.2555	.1138	**.4105**	.1876	.2366	.3001	.1890	**.1831**	**.7883**	.1757	.0800	.4427	.2630

Table 2: In-domain and Out-domain performance of DPR. The scores for out-domain denote nDCG@10. The scores over the gold query are marked in **bold**, and the better scores between queries from QGEK are underlined.

lected to convert the query's logical form into the proposed template. A query in SQ is generated from a one-to-one correspondence of KB entities, which is very similar to the form of our proposed triplet. The conversion process proceeds in the same way as mentioned in Section 3.3.

The BART-large ($d = 1024$) is fine-tuned for 5 epochs with 47,180 template-query pairs. For training the model, Adam optimizer (Kingma and Ba, 2015) is used with the batch size of 8, and the learning rate starts from 10^{-5}.

Query Generation We used the documents in the NQ train split (Kwiatkowski et al., 2019), exploited as a training dataset in DPR (Karpukhin et al., 2020), as the target dataset for query generation. The documents are converted into a template through the process described in Section 3.4. To obtain external knowledge of the subject and object, the first paragraph and category information of the Wikipedia documents are collected and inserted into the template. The generated queries and the corresponding NQ documents, input of the queries, are used in the training step of DPR.

Retriever model The dense retriever used in the training has the same structure proposed by DPR (Karpukhin et al., 2020), which has a bi-encoder structure that calculates the dot product between query and document embedding as the ranking score. The train dataset consists of the generated queries and the corresponding NQ documents for comparison with the human labeled queries of NQ. The encoder is initialized from BERT (base, uncased) (Devlin et al., 2019b). The retriever is trained with Adam optimizer (Kingma and Ba, 2015) for 25 epochs. The negative samples for contrastive learning are sampled from a single batch. The size of the train batch is 8 and the learning rate is initialized with $2 \cdot 10^{-5}$.

5 Result & Discussion

5.1 Overall Result

Our main results are shown in Table 2. We evaluate the retrieval performance of the dense retriever trained with the synthetic queries from QGEK against the gold query in the NQ train split. In the in-domain experiments, the dense retriever with the gold query of NQ showed superior performance over the retriever with QGEK. QGEK shows better performance in all metrics than the ablation case not including external knowledge in the proposed triplet. The average of NDCG@10 in out-domain experiments shows a small difference (-0.0108) between the gold queries and QGEK. In detail, the retriever trained with QGEK shows better performance on 4 datasets: DBpedia, HotpotQA, Fever, and Climate-Fever. The rest of the 9 datasets show that the retriever with the gold queries is more appropriate.

Using external knowledge gives rise to generating more appropriate queries for most datasets than not using one, though human labeled queries are more appropriate for training the dense retriever in the in-domain experiments. On the other hand, we see that QGEK gives comparable performance to the one with human labeled queries in the out-domain experiments and even outperforms on some datasets.

5.2 Analysis of Synthetic Queries

Experiments are conducted to compare against query generator baselines. We selected GenQ (Thakur et al., 2021) and Info-HCVAE (Lee et al., 2020) models as the baselines. The models receive the documents in NQ train split as input. The size

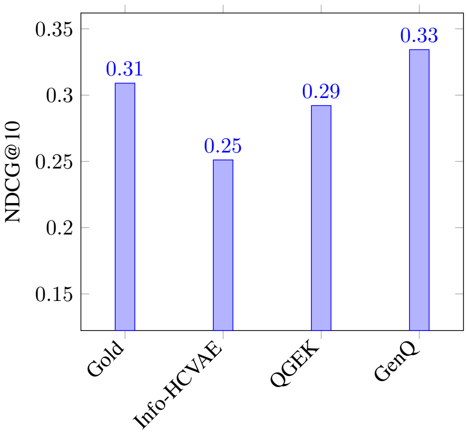

Figure 4: NDCG@10 average of the dense retrieval trained with various queries for NQ and 13 out-domain datasets.

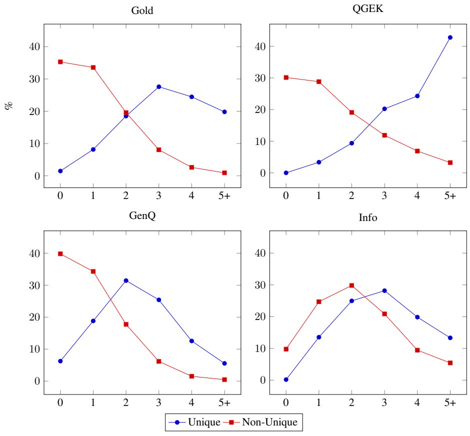

Figure 5: Distribution of unique & non-unique words in the queries.

and documents of the dataset are the same as those of the NQ train split except for synthetic queries.

Baseline Comparison A comparison with other query creation methods is made, as shown in Figure 4. The average of the NDCG@10 performance in in-domain and out-domain experiments is calculated by training the dense retriever through the generated queries. The models trained with synthetic queries are sorted as GenQ, QGEK, and Info-HCVAE in descending order. QGEK shows somewhat lower performance than the one with gold queries, but GenQ shows the best performance, indicating that many queries suitable for the IR tasks are generated by training on the MS MARCO dataset.

The MS MARCO dataset is most widely used for dense retriever training, and training a dense retriever with MS MARCO is known to give a higher performance than training it on other datasets such as NQ. Also, it has a huge amount of data, more than 500,000 pairs. This has the advantage of generating queries suitable for IR tasks based on abundant and task-appropriate data. However, the proposed method is trained on a relatively small amount of 47,180 data from SimpleQuestions, a KB-QA dataset. There is a possibility that the generated queries are largely incompatible with the IR task. However, the proposed method focuses on utilizing external knowledge, and it can be applied orthogonally to the MS MARCO dataset, which we leave for future work.

Unique & Non-Unique Words in Query We analyze whether the words in a query are from the corresponding documents. The implicitly inferring query has a higher probability of including unique words not present in the document. So, the distribution of unique & non-unique words can indirectly tell the existence of such queries. The stop words, such as the interrogative word and articles, in a query are excluded from the analysis.

The distribution of unique words in a query is shown in Figure 5. The 27% of gold labels of NQ contain 3 unique words, and 80% of the cases contain 4 or fewer unique words. QGEK shows a similar pattern of non-unique words compared with the gold, and over 40% of queries contain more than 5 unique words. The distribution of GenQ shows a similar pattern to that of the gold queries in both unique and non-unique words. Unlike other models, the Info most frequently includes 2 non-unique words.

Note that QGEK generates queries with more unique words than other queries, together with a similar distribution of non-unique words to that of gold queries. This implies that QGEK can generate queries requesting hidden information not present in the document. Given the performance of the dense retriever (Figure 4) and the distribution of unique & non-unique words (Figure 5), generating queries both close to the human labeled ones and appropriate to the IR tasks is an important factor for an optimal training of the dense retriever. Our future work includes generating queries not only close to human labeled ones but also optimized for IR tasks, such as exploiting the MS MARCO dataset.

Manual Evaluation We use human evaluation to check whether the synthetic queries are similar to human labeled ones. The randomly sampled 30

Document 1	Document 2	Document 3
(...) reporting having problems with their water treatment plants with the mussels attaching themselves to pipeworks. (...) They were first detected in Canada in the Great Lakes in 1988, in Lake St. Clair, located east/northeast of Detroit and Windsor. (...)	Call of Duty: World at War is a first-person shooter video game developed by Treyarch and published by Activision. It was released for Microsoft Windows, the PlayStation 3, Xbox 360, and Wii in November 2008. (...) "World at War" received ports featuring different storyline versions, while remaining in the World War II setting, for the and . (...)	(...) Call the Midwife is a BBC period drama series about a group of nurse midwives working in the East End of London in the late 1950s and early 1960s. It stars Jessica Raine, Miranda Hart, Helen George, Bryony Hannah, Laura Main, Jenny Agutter, Pam Ferris, (...) and Leonie Elliott. The series is produced by Neal Street Productions, a production company founded (...)
Gold Label when did zebra mussels come to **north america**	**Gold Label** who made call of duty world at war	**Gold Label** where in london is call the midwife set
QGEK What is the date zebra mussel was first detected in Canada? **(-) Ext. Knowledge** what country is zebra mussel found	**QGEK** What **console** is call of duty: world at war available on **(-) Ext. Knowledge** what is the setting of call of duty: world at war	**QGEK** who is the **actress** for call the midwife **(-) Ext. Knowledge** who produced call the midwife
Info-HCVAE where did the lake st. clairs originate?	**Info-HCVAE** what setting was the setting for the game of the " world at war :"?	**Info-HCVAE** in what time period did the bbc's the midcene series take place?
GenQ where are mussels located	**GenQ** what year did call of duty world at war come out	**GenQ** cast of call the midwife

Table 3: Examples of documents and the corresponding queries. The non-unique words are underlined, and the unique words are marked in **bold**.

documents and corresponding queries are given to three annotators fluent in English. After reading the given documents, annotators evaluated each query on a scale of 0-5 against three points: 1) how relevant a given query is to the document (Relevancy), 2) how grammatically natural it is (Grammaticality), and 3) how much reasoning is needed to answer (Difficulty).

Query	Relevancy	Grammaticality	Difficulty
Gold	3.95 (±1.38)	3.80 (±1.12)	2.10 (±1.43)
QGEK	3.66 (±1.50)	**4.07 (±1.04)**	**2.39 (±1.50)**
Info-HCVAE	3.66 (±1.45)	4.01 (±1.13)	2.31 (±1.52)
GenQ	4.12 (±1.20)	4.02 (±1.26)	1.90 (±1.21)

Table 4: The result of human evaluation. Statistically significant difference compared to gold via t-test ($p < 0.05$) is marked in **bold**.

As shown in Table 4, QGEK shows statistically higher degrees of grammaticality and difficulty than the gold labels. These results indicate that queries from QGEK need more hidden information not present in the documents compared to other queries.

Case Study Examples of the documents and corresponding queries are shown in Table 3.

Document 1 is about the water treatment problem caused by mussels. In answering the gold label, external knowledge that Canada is in North America is needed for the inference from the document. However, other generated queries do not require much external information. In the case of Document 2, the introduction of the game "Call of Duty", the gold label does not require hidden information in the document. However, in the case of GenQ, the additional information that PlayStation 3, Xbox 360, and Wii are gaming consoles is required for a suitable answer. This gives evidence that there are cases in which queries requiring inference from external knowledge are generated through the proposed method. In the case of Document 3, introduction of Call the Midwife, the query from QGEK needs external information about the gender of actors to answer.

Although QGEK generates the queries that need external knowledge to answer, they have a similar pattern that begins with an interrogative word. In the case of GenQ and Info-HCVAE, different patterns exist through the queries of Document 3. It can be inferred that the triplet-based template makes the logical structure simple, and that the syntactic diversity of the generated query tends to decrease. For future work, we plan to propose a template that can include more logical structures, developed from the current triplet-based template.

6 Conclusion

We presented a novel query generation method, QGEK, that generates synthetic queries in a form more similar to human labeled queries by using external knowledge. In order to use unprocessed external knowledge, we convert a query into a triplet-based template, which can include information of subjects and answers. Remarkably, when dense retrieval models are trained with the queries generated from QGEK, the performance has improved much compared to using the queries without external knowledge. Also, we have shown that including external knowledge give rises to the distribution of the unique words similar to that of the human labeled queries. We believe that QGEK can also be applied to the other generation methods by orthogonally adding some external knowledge processing modules. For future work, we plan to generate queries both close to human labeled ones and optimized for IR tasks and to allow the template to accept more general logical forms for diverse high-quality queries. The code and data will be made available for public access.

Acknowledgements

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00582, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection).

References

Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 192–199, New York, NY, USA. Association for Computing Machinery.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of touché 2020: Argument retrieval - extended abstract. In *CLEF*, pages 384–395.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval*, pages 716–722, Cham. Springer International Publishing.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1265–1268, New York, NY, USA. Association for Computing Machinery.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering

Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online. Association for Computational Linguistics.

Markus Leippold and Thomas Diggelmann. 2020. Climate-fever: A dataset for verification of real-world climate claims. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088, Online. Association for Computational Linguistics.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1).

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2021. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *arXiv preprint arXiv:2112.07577*.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro A. Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *EMNLP (Findings)*, pages 4129–4140.

Ji Xin, Chenyan Xiong, Ashwin Srinivasan, Ankita Sharma, Damien Jose, and Paul Bennett. 2021. Zero-shot dense retrieval with momentum adversarial domain invariant representations. *ArXiv*, abs/2110.07581.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.