Retrieval-based Question Answering with Passage Expansion using a Knowledge Graph

Benno Kruit, Yiming Xu, Jan-Christoph Kalo
Vrije Universiteit Amsterdam, Universiteit van Amsterdam
De Boelelaan 1111, 1081 HV Amsterdam, Science Park 900, 1098 XH Amsterdam
b.b.kruit@vu.nl, y13.xu@student.vu.nl, j.c.kalo@uva.nl

Abstract

Recent advancements in dense neural retrievers and language models have led to large improvements in state-of-the-art approaches to open-domain Question Answering (QA) based on retriever-reader architectures. However, issues stemming from data quality and imbalances in the use of dense embeddings have hindered performance, particularly for less common entities and facts. To tackle these problems, this study explores a multi-modal passage retrieval model's potential to bolster QA system performance. This study poses three key questions: (1) Can a distantly supervised question-relation extraction model enhance retrieval using a knowledge graph (KG), compensating for dense neural retrievers' shortcomings with rare entities? (2) How does this multi-modal approach compare to existing QA systems based on textual features? (3) Can this QA system alleviate poor performance on less common entities on common benchmarks? We devise a multi-modal retriever combining entity features and textual data, leading to improved retrieval precision in some situations, particularly for less common entities. Experiments across different datasets confirm enhanced performance for entity-centric questions, but challenges remain in handling complex generalized questions .

Keywords: Question Answering, Information Retrieval, Relation Extraction, Knowledge Graph

1. Introduction

Motivation. QA has been a rapidly developing research topic in computer science and artificial intelligence in recent years, with the goal of providing concise and accurate responses to users' queries. With the explosive growth of information on the Internet, QA systems have gained increasing importance in helping users swiftly and effortlessly locate the information they seek. This area is both useful and challenging, demanding a profound comprehension of the problem and context within data sources, which enables the extraction of pertinent information to answer the question.

In terms of model architecture, popular Open Domain QA models can be broadly classified into two main categories: retriever-reader approaches (Zhang et al., 2021; Izacard and Grave, 2021), and retriever-free approaches (Radford et al., 2019; Raffel et al., 2020). The *retriever-reader* approaches are exemplified by DrQA (Chen et al., 2017), which constructs an open-domain QA system based on Wikipedia. This methodology breaks down the Open Domain QA task into two distinct subtasks: information retrieval and reading comprehension. The information retrieval module first extracts a set of pertinent documents or passages from a vast pool of resources. Subsequently, the reader processes this collection to extract the answer within the documents and passages. The initial retriever-reader models often employed conventional sparse vector space methods for text retrieval, omitting many valuable features. Hence, information retrieval techniques based on dense representations have emerged (Lee et al., 2019; Karpukhin et al., 2020), such as Dense Passage Retrieval (DPR) and hybrid methodologies that combine dense and sparse retrieval methods (Seo et al., 2019). In this paper, we are interested in provenance-driven QA and therefore focus on the retriever-reader-based approach.

Problem. However, methods based on dense representations of documents still have limitations when dealing with uncommon entities and facts. Sciavolino et al. (2021) constructed a set of simple and entity-rich datasets called EntityQuestions, based on Wikipedia facts, and observed that dense retrievers perform significantly worse than sparse methods on uncommon entities. Mallen et al. (Mallen et al., 2023) built PopQA and found that LMs struggle with less popular factual knowledge. Using Knowledge Graphs to enhance retrieval may help address such problems.

Based on the modality of retrieval, such KG-based QA models can be categorized as structure-based QA systems (Zheng et al., 2020; Ghosh et al., 2023), and multimodal-based QA systems (Min et al., 2019; Ju et al., 2022; Yasunaga et al., 2022; Zhang et al., 2021). For instance, NQ_BERT-DM (Zheng et al., 2020) constructs a document graph to capture relationships between paragraphs, sentences, and tokens. KG-FiD (Yu et al., 2022) forms graphs based on DPR's retrieval outcomes and devises graph-based retrievers for reordering. Yet, the document-level graph features produced by these models often disregard the spe-

LREC-COLING 2024, pages 14063–14072

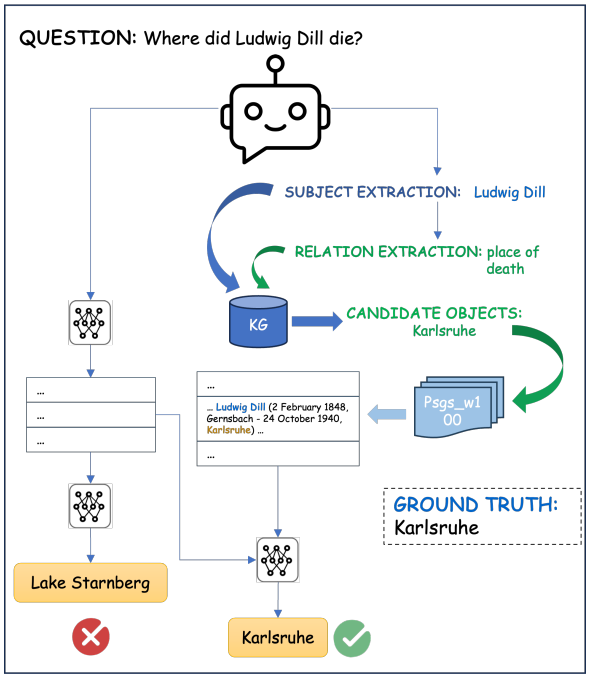


Figure 1: Overview of our proposed method for candidate passage expansion (answer on left) compared to a text-only retriever (answer on right).

cific interrelations among entities. To rectify this, certain KG-based methodologies have been introduced. GRAFT-Net (Sun et al., 2018), for instance, generates heterogeneous graphs from entities in pertinent documents and redefines question-answering as node classification within the graph. Conversely, DRAGON (Yasunaga et al., 2022) and GreaseLM (Zhang et al., 2021) aim to deeply fuse text and KGs (Section 2.3).

However, a systematic study of the strengths and limitations of using KG data for improving retriever-based QA is still missing.

Contributions. In this work, we explore the impact of multi-modal interaction-enhanced retrieval results on the performance of the QA system across different levels of entity popularity. The main contributions of this paper are as follows:

- We create a KG based on a subset of Wikidata and align it to a title-entity relationship table based on Wikipedia text. This table links wiki titles to sets of entities found hyperlinked in related paragraphs.
- We introduce a multi-modal retrieval approach that combines KGs and text features to retrieve relevant passages. These passages, combined with results from a dense retrieval model, serve as input for the reader.
- We evaluate our model on the PopQA (Mallen et al., 2023), EntityQuestions (Sciavolino et al., 2021), and NaturalQuestions datasets (Kwiatkowski et al., 2019). We binned entities into 8 levels based on popularity scores for objects and subjects. Our results highlight that our simple multi-modal interaction approach improves retriever and reader performance across both common and less popular entities.

2. Background and Related Work

In this section, we provide an overview of the retriever-reader architecture for textual retrieval-based QA, provide a short background on KGs and relation extraction, and then discuss QA approaches that combine text and KGs.

2.1. Retrieval-based Question Answering

Retriever-Reader Models. Chen et al. (2017) introduced the DrQA model, comprising a document retriever for selecting relevant paragraphs from Wikipedia articles, and a document reader module that uses them to answer prediction. The document reader follows a standard machine reading comprehension architecture, allowing for easy replacement with alternative models. Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) is a document retriever that employs dense vector representations to retrieve paragraphs matching a question from large text datasets. It calculates similarity scores using dense vectors, enhancing semantic information capture compared to traditional methods like TF-IDF. Language models (Radford et al., 2018; Kenton and Toutanova, 2019; Liu et al., 2019) have boosted QA systems by extracting rich semantic information from text. Models like BERT$_{joint}$ (Alberti et al., 2019) take questions and all retrieved passages concatenated as input for question answering as span prediction, but handling long texts can be challenging. To mitigate this, the Fusion-in-Decoder (FiD) model (Izacard and Grave, 2021) concatenates vector-encoded questions and paragraphs as input to the decoder, achieving impressive results in reading comprehension tasks.

2.2. Knowledge Graphs and Information Extraction

Knowledge Graphs. Since our approach uses a KG as an additional source of knowledge to enhance QA, we first give a short introduction into KGs and then explain how our KG was constructed.

A KG is a structured representation of interconnected facts, entities, and relationships that aims to model a domain of knowledge in a machine-readable format. KGs consist of nodes that repre-

sent entities or concepts and edges that signify the relationships or attributes connecting them. More formally a KG can be seen as a set of (s, p, o) triples, where $s, o \in E$ are entities, and $p \in R$ is from a set of relations. Imagine we want to represent the fact that "Barack Obama was born in Honolulu." In triple form, this could be represented as: Subject (s): Barack Obama Predicate (p): bornIn Object (o): Honolulu So, the triple would be (`Barack Obama`, `bornIn`, `Honolulu`).

In our project, we have used a KG based on a subset of Wikidata, a multilingual KG project with hundreds millions of data items, including descriptive information about entities such as people, places, events, and their relationships. We select a subset of "truthy" Wikidata triples for which the subject and object have an English Wikipedia article from which at least one paragraph could be extracted.[1]

Distantly Supervised Relation Extraction. To leverage the KG for answering questions, we need to identify which KG relation is expressed by a question. We frame this as a *relation extraction* task, which is usually formulated as a classification problem: given a sentence the goal is to predict one of multiple relations of a given KG schema. Large amounts of training data may be generated by *distant supervision* (Mintz et al., 2009), where training data is created automatically by associating entity pairs in sentences to the subject and object of existing KG triples. The corresponding relations are then used as gold labels for some form of fine-tuning of large pre-trained language models. While this kind of training data generation may lead to some noisy training examples, its quality is usually sufficient to produce acceptable results. In our case, the training labels (= relations / properties) are found using an entity in the question (Section 3.1) and the entity in the answer (if applicable). The relation extraction model is then only trained on questions as input (without the answer) to predict the expressed KG relation.

2.3. Text + Knowledge QA

Both text-based and structure-based approaches have excelled in QA models, leading to models like Grape (Ju et al., 2022) that integrate text features and knowledge graphs (KGs) to consider the relationships between text sequences and important entities. Specifically, this model uses the T5-Encoder for text features, while also identifying entities present in the passage and constructing a KG based on the relationships between them. Subsequently, a relation-aware GNN module fuses the features from both modalities and serve as input to the T5-Decoder, which generates the answers.

While models like KG-FiD (Yu et al., 2022) use graph-based retrieval from Dense Passage Retrieval (DPR), they may overlook specific entity interrelations. In response, some KG-based methods like GRAFT-Net (Sun et al., 2018), DRAGON (Yasunaga et al., 2022), and GreaseLM (Zhang et al., 2021) aim to deeply integrate text and KGs with specialized neural architectures. In contrast, our approach focuses on a simple yet comprehensive QA method using text and KG, addressing issues related to entities of varying popularity by conducting precise knowledge graph searches to compensate for the limitations of dense retrieval-based retrievers.

3. Approach

Fig. 1 provides an overview of our approach. Our retrieval system combines elements of DPR and multi-modal retrieval methods. When given a question, we start by utilizing ELQ to extract the subject from the question and predict the corresponding relation. Afterward, we query the knowledge graph for candidate objects related to both the subject and relation. Once we have identified the subject, relation, and candidate objects, we employ two strategies to retrieve relevant passages from textual resources. Finally, we combine the retrieved passages with DPR's retrieval results as input to the FiD reader, yielding the answer to the question.

In comparison to approaches that solely rely on DPR as the retrieval system, our model retrieves more relevant passages for less common entities, thereby assisting the reader in obtaining accurate answers.

3.1. Subject Extraction.

For EntityQuestions and NaturalQuestions, we use the state-of-the-art ELQ (Entity Linking for Questions) model (Li et al., 2020) to extract subjects from the questions. ELQ is an efficient version of BLINK (Wu et al., 2020). The model efficiently performs mention detection and linking simultaneously using a bi-encoder architecture, achieving fast and accurate entity linking for question-related tasks. Experimental results on WebQSP (Berant et al., 2013) and GraphQuestions (Su et al., 2016) demonstrate that ELQ significantly improves the performance of downstream QA tasks. In this project, we use ELQ to extract subjects from questions in EntityQuestions and NaturalQuestions datasets. We then use Wikimapper[2] to obtain the entity IDs for

[1] We use the set of English Wikipedia passages from December 2018 as extracted by Izacard and Grave (2021) to correspond to NaturalQuestions. However, the Wikidata version we use is from November 2021 due to limited accessibility of historical dumps.

[2] https://github.com/jcklie/wikimapper.

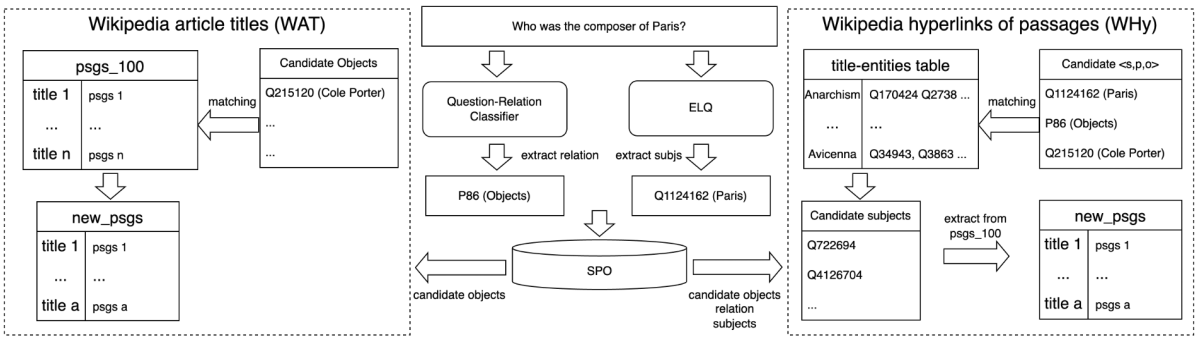


Figure 2: Details of Candidate Passage Expansion approaches

these extracted entities in Wikidata.

3.2. Relation Extraction.

For question-relation extraction, we employ a classifier that consists of a BERT model and a linear layer. Specifically, this model takes the question as input, uses the output of the BERT's CLS Token as the input to the linear layer, and trains the model through supervised learning.

During the training phase, we use a labeled dataset to perform supervised learning on the model. For each sample, the question is input to the BERT model, and the output of the CLS Token is used as the input to the linear layer. This layer performs a linear transformation on the CLS Token's output, mapping it to the number of categories required. Then, we use the cross-entropy loss function to measure the difference between the predicted results and the true labels.

Question-Relation Extraction Training Data Of the data created using distant supervision (Section 2.2), we only use items for which the label occurs more than 15 times. Table 1 presents detailed information about the training data we extracted from three benchmark datasets. Specifically, for all questions and their corresponding answers, we extract subject-object pairs. Subsequently, we match the corresponding triples that connect the subject and object from the knowledge graph to obtain accurate question-relation pairs. From this, we filter the data to include only those with more than 15 samples, which serves as the dataset for training the question-relation extraction model.

3.3. Candidate Object Lookup

We used two methods to retrieve new passages from text resources, as shown in the Fig. 2. In the Wikipedia Article Titles (WAT) approach, once candidate objects are obtained, we directly retrieve paragraphs from the passages database that have these objects as their titles. These retrieved paragraphs are then used as new passages input for the reader model. In the Wikipedia Hyperlinks of passages (WHy) approach, we take into consideration additional information. Specifically, in the preceding steps, we have acquired the subject of the question, predicted attributes based on the question, and a list of candidate objects. Using this information, we retrieve these objects from a title-entity table constructed based on the passages and perform counting. Subsequently, based on the descending order of the count results, we obtain the retrieved new paragraphs.

3.4. Final Model

We adopt the knowledge distillation-based retriever proposed by Izacard and Grave (2020). In the final set of 150 passages, the first 100 are generated by this retriever, while the remaining 50 are generated using the retriever we proposed.

We employed the fusion-in-decoder (Izacard and Grave, 2021) as the reader model, which takes each retrieved passage and the question as inputs and encodes them separately using an encoder. These encodings are then concatenated and fed into the decoder to generate the final response. This model achieved state-of-the-art performance on multiple QA datasets.

4. Experimental Setup

4.1. Benchmark Datasets

We assessed our model on three notable open benchmarks: PopQA (Mallen et al., 2023), EntityQuestions (Sciavolino et al., 2021), and NaturalQuestions (Kwiatkowski et al., 2019). PopQA and EntityQuestions are entity-centric datasets, designed to evaluate QA models on entities of varying popularity levels. In contrast, NaturalQuestions is a complex dataset tailored for open-domain QA tasks.

To analyze QA model performance on entities with different popularity levels, we extracted subjects from questions and determined the popularity

Datasets	Relations	Train	Test	Dev	RE data	RE labels	RE Acc.
PopQA	16	9,997	2,856	1,429	99.8%	16 (no DS)	97%
EntityQuestions	24	176,560	22,075	22,068	70.8%	62 / 149	79%
NaturalQuestions	(natural)	79,168	3,610	8,757	19.3%	63 / 272	74%

Table 1: Overview of datasets, and Relation Extraction classifier details, including accuracy. Only RE labels (i.e. KG relations) that were found over 15 times were used for Distant Supervision (shown as *#used* / *#found*; the PopQA model is fully supervised so all relations are used).

Datasets	Dev data				Test data			
	RE Acc.	DPR	WAT	WHy	RE Acc.	DPR	WAT	WHy
PopQA	97%	74.7	**96.2**	90.8	97%	75.8	**96.2**	90.5
EntityQuestions	79%	63.1	**82.0**	71.1	79%	63.4	**82.0**	71.2
NaturalQuestions	75%	88.7	88.7	**93.7**	74%	89.3	89.2	**92.4**

Table 2: Overall R@k in Passage Retrieval, compared to RE accuracy.

of each subject and object in the samples. Among the three datasets, PopQA already includes subject information and subject/object popularity.

For EntityQuestions and NaturalQuestions, we initially conducted entity linking, as discussed in Section 3. Subsequently, we gathered the popularity of each subject and object based on their Wikipedia page views. Specifically, we used the entity's visit count from January 1, 2022, to December 31, 2022, as a popularity measure.

PopQA is a novel large-scale open-domain QA dataset focused on entities, consisting of 14k question-answer pairs, each containing fine-grained information such as Wikidata entity IDs, Wikipedia page views, and relationship type details. The dataset is constructed by sampling knowledge triples from Wikidata and converting them into natural language questions. Specifically, the dataset involves sampling 16 different relationship types and randomly selecting knowledge triples (subject, relation, object) from Wikidata that include these relationship types. Questions are then formulated based on the subject (S) and relationship type (R) of the knowledge triple.

The *EntityQuestions* dataset comprises a collection of straightforward, entity-centric questions. Similar to the construction approach of PopQA, the authors collected triples from the T-REx dataset (Elsahar et al., 2018) and transformed them into natural language questions. The dataset contains 220k question-answer pairs, covering 24 common relationship types.

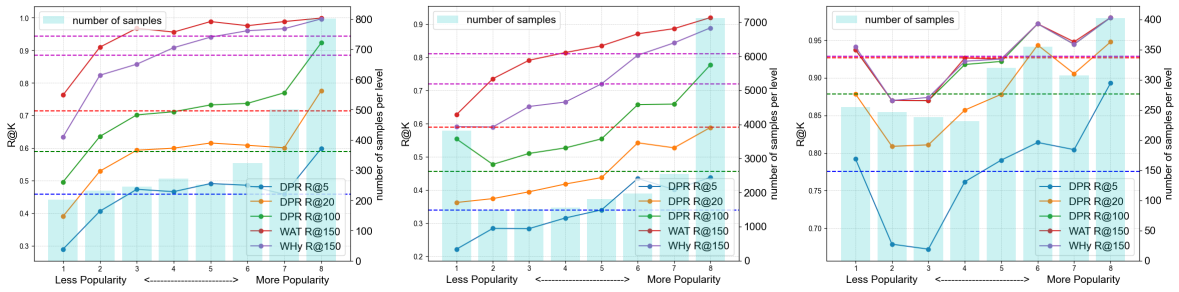
The *NaturalQuestions* corpus is a significant and comprehensive QA dataset that has made notable contributions to the advancement of natural language processing. Each example in the dataset consists of a query acquired from google.com and its corresponding Wikipedia page, which provides relevant information for answering the question. The original NaturalQuestions dataset includes both short and long answers, with approximately 1% of the questions having "yes" or "no" as the answer. For our project, we used a sampled version of the NaturalQuestions dataset, similar to the one used in FiD (Izacard and Grave, 2021). This dataset was constructed by excluding samples with answers longer than five tokens.

4.2. Metrics

We assess the model's performance in two ways: first, by evaluating its performance in the information retrieval task using the Top-k retrieval accuracy metric, and second, by evaluating its performance in the reading comprehension task using the Exact match metric.

Since the introduction of DPR (Karpukhin et al., 2020), top-k retrieval accuracy (R@k) has been widely used as a metric for measuring the performance of information retrievers. It indicates the percentage of questions for which at least one paragraph containing the ground truth answer is retrieved among the top-k paragraphs.

Exact match (EM) (Rajpurkar et al., 2016) is one of the widely used metrics in natural language processing tasks and is applied in QA systems, machine translation, information retrieval, and other tasks. In this project, we employ the exact match metric to assess the accuracy of our model's answers. The exact match metric determines whether the model's answer exactly matches the correct answer. Specifically, for each question, if the model's answer matches any of the standard answers, it is counted as 1; otherwise, it is counted as 0. The accuracy is then computed for the entire set of questions based on these counts.



(a) PopQA (b) EntityQuestions (c) NaturalQuestions

Figure 3: Retriever performance (R@k) per level of subject entity popularity . In most settings, the retriever is better at finding the correct passage for entity-centric questions and answers containing popular entities.

5. Results

5.1. Relation Extraction Classifier

In Table 1, we report statistics on the benchmarks and the performance of the RE classifier. The model trained on PopQA (without distant supervision) is able to attain very high accuracy, because it can simply learn the question templates that were used per relation for the creation of the dataset. On EntityQuestions, most data can be used for distant supervision. However, the use of entity linking for finding the subject entity creates noise, which impacts the performance. The level of this noise is highlighted by the fact that the model is trained on 62 identified relations, instead of the 24 relations that were used in the creation of the dataset. On NaturalQuestions, less than 20% of the data can be used for distant supervision. This is due to the fact that this dataset is not entity-centric, nor artificially created from a KG. Again, the large number of (possibly spurious) relations impacts the model performance.

5.2. Quantitative Results

Retriever only Table 2 presents the overall performance of the retrieval system across the entire dataset, while Fig. 3 display its performance on subjects with different popularity levels (equal-sized bins). When considering the overall data (Table 2), it is evident that DPR+WAT significantly improves performance on PopQA and EntityQuestions. This indicates that for entity-centric datasets, our relation extraction classifier model successfully predicts the corresponding relations for questions and extracts the correct candidate objects from the KG.

From 3, we observe that the passage retrieval performance improves with higher subject and object popularity, especially for entity-centric question benchmarks. The WAT passage retrieval apprach clearly approach outperforms WHy, implying that paragraphs from articles about the object are more useful than paragraphs where the object is mentioned. Note that in general, there is an imbalance in the data regarding popular entities, which occur in more questions than impopular entities (depicted by the blue histogram bars).

Retriever + Reader From Table 3, we observe that KG-enhanced candidate passage expansion can improve retrieval-based QA on entity-centric questions. Especially the WAT approach results in large improvements on the Exact Match (EM) scores of the full model. On NaturalQuestions, however, the improvement is minimal. This may be due to several factors, which are discussed in the next section (Section 5.3).

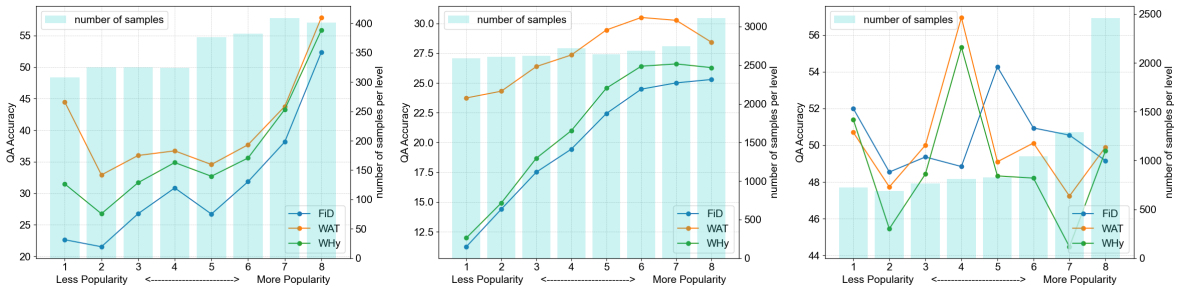
From Fig. 4, we observe that the open-domain QA performance (retrieval + reading) improves with higher subject popularity on entity-centric question benchmarks, especially using WAT. Again, WAT outperforms WHy, implying that paragraphs from articles about the object are more useful than paragraphs where the object is mentioned.

5.3. Error Analysis

From experimental results, it is evident that the retriever-only and retriever + reader models perform admirably on entity-centric datasets. However, their performance improvements are somewhat limited on NaturalQuestion datasets. We posit that this limitation may stem from the inherent difficulty of effectively extracting subject and object entities from NaturalQuestions. This challenge hampers our ability to construct a robust question-relation extraction model suitable for such questions and consequently hinders our ability to obtain effective

	Dev dataset			Test dataset		
	PQA	EQ	NQ	PQA	EQ	NQ
DPR	32.2	20.0	48.4	31.3	20.9	50.1
WAT	**41.1**	**26.1**	**49.1**	**40.5**	**26.2**	**50.2**
WHy	37.6	21.5	48.3	36.5	22.7	48.9

Table 3: Full pipeline results: retriever (indicated) + reader (FiD) performance, as Exact Match (EM) scores



(a) PopQA (b) EntityQuestions (c) NaturalQuestions

Figure 4: Exact Match (EM) performance of retrieval + reader per level of subject entity popularity .

candidate objects, thus restricting the model's performance on this class of questions.

To substantiate this hypothesis, in this chapter, we will conduct an analysis of the effectiveness of object retrieval results across three benchmark datasets and delve into interesting errors of certain adverse experimental outcomes.

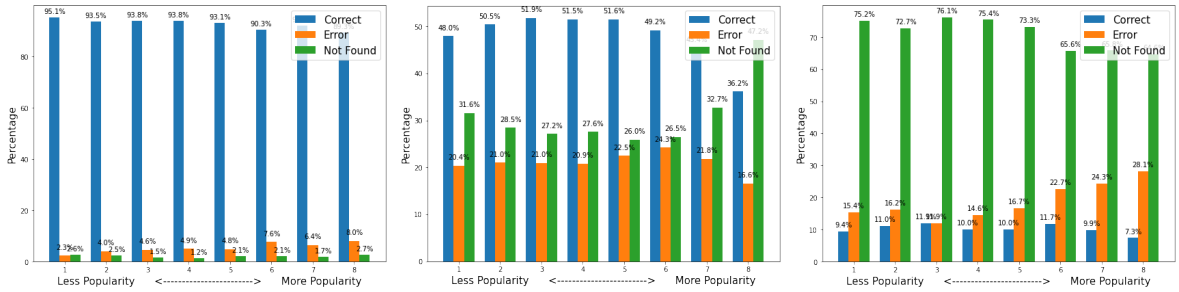
Object Retrieval Fig. 5 illustrate the accuracy of candidate object retrieval by the model for questions at different levels of subject popularity across three datasets. Candidate objects are categorized into three cases: Correct, Error, and Not Found. In this context, Correct refers to candidate objects that match the candidate answers, "error" indicates candidate objects that do not match the candidate answers, and "not found" signifies instances where the model fails due to the inability to retrieve corresponding objects from the knowledge graph based on predicted relations and question subjects. It is evident that on the PopQA dataset, the majority of retrieved objects align with the candidate answers. This is attributed to the dataset's provision of explicit subject and object information, enabling us to construct effective question-relation extraction models and thereby enabling the model to accurately retrieve relevant objects. However, on the EntityQuestions dataset, while retrieved objects are mostly correct across most levels of popularity, the proportions of error and not-found objects increase. This could be attributed to limitations in the subject extraction method based on ELQ, preventing us from identifying the correct subjects to support the construction of robust question-relation extraction models.

In the case of the NaturalQuestions dataset, it can be observed that the quality of retrieved candidate objects is quite poor. Across all levels, a significant proportion falls under the Not Found category, indicating subpar subject extraction and relation prediction. This stems from the characteristics of the NaturalQuestions dataset, which includes complex and generalized questions that ELQ struggles to accurately identify subjects for. Moreover, many questions involve not just entities but also sentences composed of phrases. In such scenarios, our model fails to retrieve relevant objects from the knowledge graph and obtain effective passages, thus hindering improvements to the baseline model's performance.

Examples Analysis Table 4 illustrates several limitations of the model. In examples 1 and 2, the baseline model could originally provide the correct answers, and our model accurately retrieved candidate objects. However, introducing paragraphs retrieved by the multimodal retriever added some noise, resulting in the reader providing incorrect answers. In examples 3 and 4, our model identified the correct candidate objects but did not change the baseline's incorrect answers. We believe that further fine-tuning of the reader using the results from DPR and the multimodal retriever may address these limitations. Examples 5 and 6 highlight the limitations of ELQ, where the subject extraction model failed to correctly extract subjects from the questions, rendering the multimodal retriever ineffective. In example 7, although the subject extraction successfully extracted the correct subjects, the relation prediction model failed to predict the question correctly, causing the model to be unable to obtain the correct candidate objects and paragraphs. Examples 8 and 9 demonstrate issues present in the NaturalQuestions dataset, such as generalized questions and answers containing complex non-entity information.

6. Conclusion

In this work, we propose a QA system that integrates text and KGs. Through comparisons with robust benchmark models, we investigate the impact of a multimodal QA system on questions related to less common entities. We validate our model on two entity-centric datasets, PopQA and EntityQuestions, as well as on a complex general question dataset, NaturalQuestions, and compare it with the FiD model serving as the baseline. Overall, our model outperforms the baseline across the entire dataset. In entity-centric QA datasets, our



(a) PopQA (b) EntityQuestions (c) NaturalQuestions

Figure 5: Object retrieval performance from KG, per level of subject popularity

	Question	Correct Answer	Baseline Answer	Triple	Our Answer	Error Type
1	What music label is You Got My Attention represented by?	Fervent Records	Fervent Records ✓	**s:** You Got My Attention **r:** record label **o:** Fervent Records ✓	eMusic ✗	noisy
2	Who is the author of Before?	Gael Baudino	Gael Baudino ✓	**s:** Before **r:** author **o:** Gael Baudino ✓	Franz Kafka ✗	noisy
3	Who is the author of Black Orchids?	Rex Stout Rex Todhunter Stout	Neil Gaiman ✗	**s:** Black Orchids **r:** author **o:** Rex Stout ✓	Neil Gaiman ✗	weak reader
4	Who sings I want to dance with you?	George Strait	Baby K ✗	**s:** I Just Want to Dance with You **r:** performer **o:** George Strait ✓	Baby K ✗	weak reader
5	Where did the battle of bonhomme richard take place ?	near Flamborough Head	the Isle of Wight ✗	**s:** Battle of Damme ✗ **r:** location **o:** Damme ✗	France ✗	wrong subject
6	Who sings love me tender in princess diaries 2?	Norah Jones Adam Levy	Julie Andrews ✗	**s:** Prince + Princess 2 ✗ **r:** cast member **o:** Dylan Kuo ✗	Julie Andrews ✗	wrong subject
7	When did i'm like a bird come out? (NaturalQuestions)	October 24, 2000	2001 ✗	**s:** I'm Like a Bird **r:** performer ✗ **o:** Nelly Furtado ✗	2001 ✗	wrong relation
8	Where is wake forest located in north carolina? (NaturalQuestions)	78.51889°W, in Franklin and Wake counties, 35.97333°N	Wake Forest, North Carolina ✗	**s:** Wake Forest, North Carolina **r:** located in the administrative territorial entity **o:** Wake County, North Carolina	Wake Forest ✗	complex answer
9	When was the first book made into a movie? (NaturalQuestions)	1924	1952 ✗	**s:** When Were You Born ✗ **r:** cast member ✗ **o:** Margaret Lindsay, Anna May Wong ✗	1952 ✗	general question

Table 4: Interesting errors of our approach

multimodal QA system compensates for the underperformance of the FiD model when dealing with less common entities. This suggests that, in contrast to DPR models unable to accurately learn representations for uncommon entities, the precise retriever based on entity and KG interaction efficiently retrieves candidate passages for less popular entities. However, in the NaturalQuestions dataset, the performance improvement of our multimodal QA system over the FiD model is limited. This is due to the insufficient training data available for the question-relation classifier module in the multimodal retriever, causing it to fail in predicting candidate relations and consequently, to leverage the KG and text resources to retrieve highly relevant passages. The research presented in this work holds significant importance as it addresses issues present in current QA systems: Can QA models combining multimodal features address the underperformance of dense neural retrieval and language model-based QA systems when dealing with less common entities and facts? Despite numerous QA models integrating multimodal features proposed in past studies, no research has explored whether multimodal models can enhance the performance of QA systems on less common entities in comparison to text-based models. The results of this study demonstrate that multimodal QA systems can enhance accuracy for less common entities, narrowing the performance gap between less and more popular entities.

Nevertheless, our model still exhibits limitations that need improvement in future work. Despite focusing on enhancing the retriever using multimodal features to mitigate performance disparities between entities of varying popularity, the improvement is more prominent in subjects than in objects. Developing a reader enhanced with multimodal features might address this issue. On the EntityQuestions dataset, our model's performance is unsatisfactory. Extracting more accurate subjects from these complex questions could aid in constructing a better question-relation classifier, thereby achieving enhanced QA system performance through rele-

vant passage retrieval. Additionally, due to computational resource and time limitations, we only applied shallow interactions between multimodal data to enhance the retriever. In future work, exploring deep interactions between various modal features could yield a more robust QA model.

7. References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

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

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Shrestha Ghosh, Simon Razniewski, and Gerhard Weikum. 2023. Answering count questions with structured answers from text. *Journal of Web Semantics*, 76:100769.

Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*.

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

Mingxuan Ju, W. Yu, Tong Zhao, Chuxu Zhang, and Yanfang Ye. 2022. Grape: Knowledge graph enhanced passage reader for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*.

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6086–6096. Association for Computational Linguistics.

Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. In *EMNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Knowledge guided text retrieval and reading for open domain question answering. *ArXiv*, abs/1911.03868.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. (August):1003–1011.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Conference on Empirical Methods in Natural Language Processing*.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4430–4441, Florence, Italy. Association for Computational Linguistics.

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gür, Zenghui Yan, and Xifeng Yan. 2016. On generating characteristic-rich question sets for qa evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 562–572.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2022. KG-FiD: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4961–4974, Dublin, Ireland. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2021. Greaselm: Graph reasoning enhanced language models. In *International conference on learning representations*.

Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu. 2020. Document modeling with graph attention networks for multi-grained machine reading comprehension. pages 6708–6718.