DuReader$_{\text{retrieval}}$: A Large-scale Chinese Benchmark for Passage Retrieval from Web Search Engine

Yifu Qiu[1]†, **Hongyu Li**[2], **Yingqi Qu**[2], **Ying Chen**[2], **Qiaoqiao She**[2],
Jing Liu[2]*, **Hua Wu**[2], **Haifeng Wang**[2]

[1]Institute for Language, Cognition and Computation, University of Edinburgh, UK
[2]Baidu Inc., Beijing, China
y.qiu-20@sms.ed.ac.uk
{lihongyu04, quyingqi, chenying04, sheqiaoqiao, liujing46, wu_hua, wanghaifeng}@baidu.com

Abstract

In this paper, we present DuReader$_{\text{retrieval}}$, a large-scale Chinese dataset for passage retrieval. DuReader$_{\text{retrieval}}$ contains more than 90K queries and over 8M unique passages from a commercial search engine. To alleviate the shortcomings of other datasets and ensure the quality of our benchmark, we (1) reduce the false negatives in development and test sets by manually annotating results pooled from multiple retrievers, and (2) remove the training queries that are semantically similar to the development and testing queries. Additionally, we provide two out-of-domain testing sets for cross-domain evaluation, as well as a set of human translated queries for for cross-lingual retrieval evaluation. The experiments demonstrate that DuReader$_{\text{retrieval}}$ is challenging and a number of problems remain unsolved, such as the salient phrase mismatch and the syntactic mismatch between queries and paragraphs. These experiments also show that dense retrievers do not generalize well across domains, and cross-lingual retrieval is essentially challenging. DuReader$_{\text{retrieval}}$ is publicly available at `https://github.com/baidu/DuReader/tree/master/DuReader-Retrieval`.

1 Introduction

Passage retrieval requires systems to select candidate passages from a large passage collection. In recent years, pre-trained language models (Devlin et al., 2019; Liu et al., 2019) have been applied to retrieval problems, known as *dense retrieval* (Karpukhin et al., 2020; Qu et al., 2021; Zhan et al., 2021). The success of dense retrieval relies on the availability of high quality, large-scale, human-annotated corpora. A number of popular datasets are already available for English passage retrieval, including MS-MARCO (Nguyen et al., 2016), TriviaQA (Joshi et al., 2017), and Natural Questions (Kwiatkowski et al., 2019). In contrast, existing datasets for non-English retrieval (e.g., Chinese), are either small or machine generated. For example, TianGong-PDR (Wu et al., 2019) has only 70 questions and 11K passages. Even though the multilingual dataset mMARCO(Bonifacio et al., 2021) is large in size, it is constructed by machine translation from the English MS-MARCO dataset. Sougou-QCL (Zheng et al., 2018) is constructed based on click logs of web data without human annotation. In this paper, we present DuReader$_{\text{retrieval}}$, a large-scale Chinese dataset for passage retrieval from web search engine, that is manually annotated. The dataset contains more than 90K queries and over 8M unique passages. All queries are selected from real requests made by users at Baidu Search, and document passages are from the search results. Similar to (Karpukhin et al., 2020), we create the DuReader$_{\text{retrieval}}$ from DuReader (He et al., 2018), a Chinese machine reading comprehension dataset, and obtain the human labels for paragraphs by distant supervision (See Section 2.2). An example from DuReader$_{\text{retrieval}}$ is shown in Table 1, and a comparison of different datasets is shown in Table 2.

Additionally, recent works point out two major shortcomings of the development and testing sets in the existing datasets:

- Arabzadeh et al. (2021) and Qu et al. (2021) observe that false negatives (i.e. relevant passages but labeled as negatives) are common in the passage retrieval datasets due to their large scale but limited human annotation. As a result, the top passages retrieved by models may be superior to labeled relevant positives, and this will affect the evaluation.
- Lewis et al. (2021) find that 30% of the test-set queries in the common machine reading comprehension datasets (Kwiatkowski et al., 2019; Joshi et al., 2017) have a near-duplicate para-

[†] The work was done when the first author was doing internship at Baidu.

[*] Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5326–5338
December 7-11, 2022 ©2022 Association for Computational Linguistics

Query:
太阳花怎么养
How to raise Grandiflora?

Positive Psg. 1 :
百度经验:jingyan.baidu.com花卉名称:太阳花播种时间:春、夏、秋均可播种为一年生肉质草本植物。株高10～15cm。花瓣颜色鲜艳,有白、深黄、红、紫等色。园艺品种很多,有单瓣、半重瓣、重瓣之分。喜温暖、阳光充足而干燥的环境,极耐瘠薄,一般土壤均能适应,能自播繁衍。见阳光花开,早、晚、阴天闭合,故有太阳花、午时花之名。花期6～7月。太阳花种子非常细小。常采用育苗盘播种,极轻微地覆些细粒蛭石,或仅在播种后略压实,以保证足够的湿润。发芽温度21～24℃,约7～10天出苗,幼苗极其细弱,因此如保持较高的温度,小苗生长很快,便能形成较为粗壮、肉质的枝叶。这时小苗可以直接上盆,采用10厘米左右直径的盆,每盆种植2～5株,成活率高,生长迅速。
Baidu experience: jingyan.baidu.com Flower name: Grandiflora Sowing time: Spring, summer, and autumn can be sown as an annual succulent herb. Plant height is 10-15cm. The petals are bright in color, white, dark yellow, red, purple and other colors. There are many horticultural varieties, including single, semi-double and double petals. It likes a warm, sunny and dry environment, is extremely tolerant to barrenness, and can adapt to general soils and reproduce by itself. It is named Grandiflora because it blooms when the sun is rising and closes in the morning, evening and cloudy days. It flowers from June to July. Grandiflora's seeds are very small. The seedling trays are often used for sowing, very lightly covered with fine vermiculite, or only slightly compacted after sowing to ensure sufficient moisture. The germination temperature is 21~24°C and the seedlings emerge in about 7~10 days. The seedlings are extremely thin. Therefore, if the temperature is kept high, the seedlings will grow quickly, and thicker, fleshy branches and leaves can be formed. At this time, the seedlings can be directly put into pots, using pots with a diameter of about 10 cm, planting 2 to 5 plants per pot, with high survival rate and rapid growth.

Positive Psg. 2 :
抹平容器中培养土平面,将剪来的太阳花嫩枝头插入竹筷戳成的洞中,深入培养土最多不超过2厘米。为使盆花尽快成形、丰满,一盆中可视花盆大小,只要能保持2厘米的间距,可扦插多株(到成苗拥挤时,可分栽他盆)。接着浇足水即可。新扦插苗可遮阴,也可不遮阴,只要保持一定湿度,一般10天至15天即可成活,进入正常的养护。太阳花极少病虫害。平时保持一定湿度,半月施一次千分之一的磷酸二氢钾,就能达到花大色艳、花开不断的目的。如果一盆中扦插多个品种,各色花齐开一盆,欣赏价值更高。每年霜降节气后(上海地区)将重瓣的太阳花移至室内照到阳光处。入冬后放在玻璃窗内侧,让盆土偏干一点,就能安全越冬。次年清明后,可将花盆置于窗外,如遇寒流来袭,还需入窗内养护。
Flatten the soil surface in the container, insert the cut branches of Grandiflora into the hole made by the bamboo chopsticks, and deepen the soil for no more than 2 cm. To make the potted flowers take shape and fullness as soon as possible, multiple plants can be cut as long as the spacing of 2 cm can be maintained (when the seedlings are crowded, they can be planted in other pots). Then pour plenty of water. The new cuttings can be shaded or not. As long as they maintain a certain humidity, they can survive 10 to 15 days and enter normal maintenance. Grandiflora has very few pests and diseases. Maintain a certain humidity at ordinary times, and apply one-thousandth of potassium dihydrogen phosphate once a half month to achieve the purpose of large flowers and continuous blooming. If there are multiple varieties of cuttings in one pot, the flowers of all colours will bloom in one pot, and the appreciation value will be higher. Every year after the frost falls (Shanghai area), the double-flowered Grandiflora is moved indoors to shine in the sun. Put it on the inside of the glass window after the winter, and let the potting soil dry a little to survive the winter safely. After the Qingming Festival in the following year, the flowerpots can be placed outside the window.

Table 1: A data instance randomly selected from the DuReader$_{\mathbf{retrieval}}$ development set.

phrase in their corresponding training sets, thus leaking the testing information into model's training. The similar issue has been observed in MS-MARCO (Zhan et al., 2022).

In the construction of DuReader$_{\mathbf{retrieval}}$, we try to alleviate the above issues and improve the quality of the development and testing sets in the following two ways (see Section 2.3):

- To reduce the false negatives in the development and testing set, we invite the internal data team to manually check and relabel the passages in the top retrieved results pooled from multiple retrievers.
- To reduce the leakage of testing information into model's training, we use a query matching model from (Zhu et al., 2021) to identify and remove the training queries that are semantically similar to the development and testing queries.

Moreover, inspired by Thakur et al. (2021), we provide two testing sets (see Section 2.4) from the medical domain (cMedQA (Zhang et al., 2018) [*] and cCOVID-News[†]) as the separate testing sets for out-of-domain evaluation. Additionally, we also provide a set that contains human translated queries for cross-lingual retrieval evaluation (see Section 2.5) (Asai et al., 2021a; Sun and Duh, 2020).

In this paper, we conduct extensive experiments. In our in-domain experiments, we find that there are many challenges to be addressed, such as salient phrase mismatches and syntactic mismatches (see Section 3.5). It is also difficult for dense retrievers to generalize well across different domains as we observed in the out-of-domain experiments (see Section 3.6). Finally, the cross-lingual experiments indicate that cross-lingual retrieval is essentially challenging (see Section 3.7).

We summarize the characteristics of our dataset and our contributions as follows:

- We present a large-scale Chinese dataset for benchmarking the passage retrieval models. Our dataset comprises more than 90K queries and more than 8M unique passages from Baidu Search.
- We leverage two strategies to improve the quality of our benchmark and alleviate the existing shortcomings in other existing datasets, including reducing the false negatives with human annotations on pooled retrieval results, and remove the training queries semantically similar to the development and testing queries.
- We introduce two extra out-of-domain test sets to evaluate the domain generalization capability, and a cross-lingual set to assess the cross-lingual retrievers.
- We conduct extensive experiments and the results demonstrate that the dataset is challenging and passage retrieval has plenty of room for improvement.

2 DuReader$_{\mathbf{retrieval}}$

In this section, we introduce our DuReader$_{\mathbf{retrieval}}$ dataset (See dataset statistics in Table 3). We first formally define the passage retrieval task in Section

[*] avaliable at https://github.com/zhangsheng93/cMedQA2
[†] available at https://www.datafountain.cn/competitions/424

Dataset	Lang	#Que.	#Psg.	Source of Que.	Source of Psg.	Psg. Annotation
MS-MARCO (Nguyen et al., 2016)	EN	516K	8.8M	User logs	Web doc.	Human
TriviaQA (Joshi et al., 2017)	EN	95K	650K	Trivia web.	Wiki./Web doc.	Dist. Sup.
Natural Questions (Kwiatkowski et al., 2019)	EN	61K	21M	User logs	Wiki doc.	Dist. Sup.
mMARCO-Chinese (Bonifacio et al., 2021)	CN	516K	8.8M	User logs	Web doc.	Tranlation
cCOVID-News[‡]	CN	4.9K	5K	User question	COVID-19 News doc.	Human
cMedQA-2.0 (Zhang et al., 2018)	CN	108K	203K	User question	Medical Forum	Question-Answer Pairs
TianGong-PDR (Wu et al., 2019)	CN	70	11K	User logs	News doc.	Human
Sougou-QCL (Zheng et al., 2018)	CN	537K	9M	Click logs	Web data	Click signal
DuReader$_{\mathbf{retrieval}}$ (Our work)	**CN**	**97K**	**8.9M**	**User logs**	**Web doc.**	**Dist. Sup. + Human**

Table 2: Summary of data statistics for passage retrieval datasets. The annotation of passages in TriviaQA and Natural Questions are presented in (Karpukhin et al., 2020). Compared with other works, the instances in DuReader$_{\mathbf{retrieval}}$ come from user logs in web search. Its consists of a distant supervised (Dist. Sup.) training set and human-annotated (Human) development and test sets.

2.1. We then introduce how we initially construct our dataset from DuReader by distant supervision in Section 2.2. Our strategies for further improving the data quality are discussed in Section 2.3. Finally, we introduce two out-of-domain test sets in Section 2.4 and a cross-lingual set in in Section 2.5.

2.1 Task Definition

DuReader$_{\mathbf{retrieval}}$ is created for the task of passage retrieval, that is, retrieving a list of relevant passages in response to a query. Formally, given a query q and a large passage collection $\mathcal{P}$, a retrieval system $\mathcal{F}$ is required to return the top-K relevant passages $P_K^{(q)} = \left\{p_1^{(q)}, p_2^{(q)}, ..., p_K^{(q)}\right\}$, where K is a manually defined number. Ideally, all the relevant passages to q within $\mathcal{P}$ should be included and ranked as high as possible in the retrieved results $P_K^{(q)}$.

2.2 Dataset Construction

2.2.1 An Introduction of DuReader Dataset

DuReader$_{\mathbf{retrieval}}$ is developed based on the Chinese machine reading comprehension dataset DuReader (He et al., 2018). All queries in DuReader are posed by the users of our chosen commercial search engine, and document-level contexts are gathered from search results. Each instance in DuReader is a tuple $< q, t, D, A >$, where q is a query, t is a query type, D is the top-5 retrieved documents constituted by their paragraphs returned by our chosen commercial search engine. A is the answers written by human annotators.

DuReader$_{\mathbf{retrieval}}$	**Train**	**Dev.**	**Test**
#Chinese queries	86,395	2,000	4,000
#passages	222,395	9,863	19,601
#avg. passages per query	2.57	4.93	4.90
#avg. Chinese characters per query	9.51	9.29	9.23
#avg. Chinese characters per passage	358.58	398.59	401.61
cMedQA (Out-of-domain)	**Train**	**Dev.**	**Test**
#queries	\	\	3,999
#passages	\	\	7,527
#avg. passages per query	\	\	1.88
#avg. Chinese characters per query	\	\	48.47
#avg. Chinese characters per passage	\	\	100.58
cCOVID-News (Out-of-domain)	**Train**	**Dev.**	**Test**
#queries	\	\	949
#passages	\	\	964
#avg. passages per query	\	\	1.02
#avg. Chinese characters per query	\	\	25.93
#avg. Chinese characters per passage	\	\	1430.93
Translated queries for cross-lingual retrieval			
#avg. English words per query	6.41	6.55	6.46
#English queries	9,500	2,000	4,000
Size of the total paragraph collection		8,096,668	

Table 3: Summary of statistics for the training (Train), development (Dev.), testing (Test), out-of-domain (OOD) testing sets and cross-lingual set of DuReader$_{\mathbf{retrieval}}$.

2.2.2 Constructing DuReader$_{\mathbf{retrieval}}$ from DuReader

In this section, we describe that how we construct DuReader$_{\mathbf{retrieval}}$ from DuReader. First, we describe our approach to labelling the positive passages. Then, we discuss our approaches to dealing with the two challenges in constructing DuReader$_{\mathbf{retrieval}}$ from DuReader: 1) the original paragraphs are too short to provide meaningful context; and 2) the term overlap between the queries and the document titles may ease the challenges for passage retrieval.

Distant Supervision for Annotations Following MS-MARCO Passage Ranking (Nguyen et al., 2016), we use the human-written answers from DuReader (He et al., 2018) to label the positive passages by the distant supervision. A paragraph is considered positive if it contains any human-written answer. Specifically, we leverage the span-level F1 score to measure the match between each human-written answer and the paragraphs in documents. If a span-answer pair gets a F1-score higher than the threshold (0.5), we label the paragraph as positive. We show the details of our annotation process in Algorithm 1.

Algorithm 1 Span-level F1 Annotation for Positives

Input: $\{\langle p, a\rangle\}$, p: candidate paragraph, a: answer, τ: threshold for positive labelling.

Output: $l_p \in \{0, 1\}$: label, 0 and 1 denote negative and positive p , separately.

```
for any span $s$ in $p$ do
    if Calculate $F1(s, a) \geq \tau$ then:
        $l_p \leftarrow 1$
        return
    end if
end for
$l_p \leftarrow 0$
return
```

Passage Length Control Additionally, most paragraphs in DuReader are too short to form meaningful contexts. We concatenate the paragraphs of each document in DuReader by the following rules: 1) In a document of less than 256 Chinese characters, all paragraphs are concatenated into one passage; 2) In a document of more than 256 Chinese characters, a paragraph of less than 256 is concatenated with the next one, and the concatenation does not stop until the length of the new passage exceeds 256. The new passage is labelled as positive if any of its components are originally labelled positive in DuReader. After the processing, the median and the mean of the passage length are 304 and 272, respectively.

Removing Document Titles We remove the titles from all documents in DuReader, since we observe that there is many term overlaps between the queries and the titles. If we keep them, the retrieval systems may easily match the queries with the document titles and achieve high performance. But we expect the retrievers to capture all contextual information in passages to answer queries.

2.3 Quality Improvement

As we discussed in the previous section, there are shortcomings of other existing datasets. To alleviate such shortcomings, we further design two strategies to ensure the quality of the development and test sets in DuReader$_{\texttt{retrieval}}$. Although in this work we apply our quality improvement approaches to the Chinese passage retrieval dataset, the proposed method allows flexibility extended to other languages (e.g., English), benefiting the future evaluation and development of dense retrieval systems.

Reducing False Negatives A common issue in existing passage retrieval datasets (Qu et al., 2021; Arabzadeh et al., 2021) is false negatives, i.e., query-relevant passages not labelled as positives, in the development and testing sets. In this section, we discuss our strategy for reducing the false negatives in the development and testing sets of DuReader$_{\texttt{retrieval}}$.

We use human annotation as a complement to the distant supervised labeling approach discussed in Section 2.2. We invite the internal data team to manually check the labels in the development and test sets and fix them if necessary. To avoid inductive bias in our annotation process, we follow the pooling method in TREC competitions (Voorhees et al., 2005) to select candidate passages for annotation. The top-ranked passages retrieved for each query by a set of *contributing retrievers* are pooled for annotation. In particular, the annotator is presented with a query and the top-5 passages pooled from five retrieval systems. We use BM25 and four neural retrievers with the initialization from ERNIE (Sun et al., 2019), BERT (Devlin et al., 2019; Cui et al., 2021), RoBERTa (Liu et al., 2019) and MacBERT (Cui et al., 2020) to serve as our contributing retrievers. We combine their top-50 retrieved passages as candidates. An ensembled re-ranker is then used (See Appendix A.1 for implementation details) to select the top-5 passages for human annotation. To ensure data quality, we perform all annotations on our internal annotation platform. Please refer to the Appendix A.3 for annotation settings and quality control.

After adopting our strategy for reducing false negatives, the average positive paragraph per query has increased from 2.43 to 4.91. 71.53% of queries have at least one false negative relabeled by annotators, which shows there are many false negatives in the raw corpus derived directly from DuReader.

Removing Similar Queries Retrieval systems should avoid merely memorizing queries and their relevant items in the training set and directly applying such memorization during inference. Lewis et al. (2021) find that in some popular datasets, including Natural Questions (Kwiatkowski et al., 2019), WebQuestions (Berant et al., 2013) and TriviaQA (Joshi et al., 2017), 30% of the test-set queries have a near-duplicate paraphrase in their corresponding training sets, which leaks the testing information into the model training. In this paper, we use a model-based approach to remove training queries that are semantically similar to development and testing queries.

We use the query matching model in (Zhu et al., 2021), which computes the similarity score ranging between $[0, 1]$ for a query pair. We set a threshold of 0.5, meaning that if the similarity between a training query and a test query is higher than 0.5, we mark the query pair as semantically similar. There are 566 training queries semantically similar to 387 queries in the development and the test set, accounting for approximately 6.45% of total development and test queries. All these 566 training instances are removed in DuReader$_{\text{retrieval}}$.

2.4 Out-of-domain Evaluation

Recent work (Thakur et al., 2021) reveals that the dense retrievers do not generalize well cross-domain. To assessing the cross-domain generalization ability of the retrievers, we carefully select two publicly available Chinese text retrieval datasets, i.e., cMedQA (Zhang et al., 2018) created from online medical consultation text and cCOVID-News from COVID-19 news articles. We randomly select 949 and 3,999 samples from cCOVID-News and cMedQA, respectively, as out-of-domain testing data.

2.5 Cross-lingual Evaluation

Cross-lingual passage retrieval has recently received much attention (Shi et al., 2021; Asai et al., 2021b), which aims to retrieve the passages in the target language (e.g., Chinese) as the response to the query in source language (e.g., English).

In DuReader$_{\text{retrieval}}$, we provide a cross-lingual retrieval set which contains the English queries paired with Chinese positive passages. The total numbers of training/development/testing English queries are 9.5K/4K/2K, respectively. All English queries in our cross-lingual set are translated and the passage annotations are aligned with DuReader$_{\text{retrieval}}$. To obtain English queries, we first translate Chinese queries to English queries by using machine translation [§]. Then, we ask the internal data team to manually check the quality of the machine-translated queries, and modify the translated queries if necessary. The quality controls for translated queries are the same as our previous human annotations for the in-domain development and testing set as in Appendix A.3.

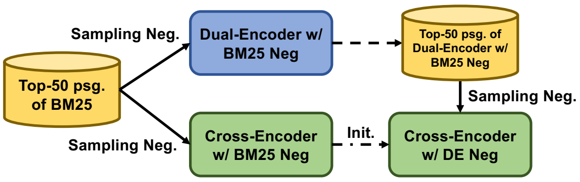

Figure 1: Illustration for the training procedure of our one dual-encoder retriever and two cross-encoder re-rankers. We train our first retriever and re-ranker by the negatives sampled from BM25's output as in (Karpukhin et al., 2020). We further attempt the strategy in (Xiong et al., 2021) that sampling negatives from dual-encoder retriever to enhance the cross-encoder re-ranker.

3 Experiments and Results

3.1 Baselines

We use the recent two-stage framework (retrieve-then-rerank) (Dang et al., 2013; Qu et al., 2021) for passage retrieval and evaluate two retrieval and two reranking models on our DuReader$_{\text{retrieval}}$ dataset. In particular, we utilize the dual-encoder and cross-encoder architecture in RocketQA (Qu et al., 2021) to develop our neural retrievers and re-rankers. We introduce the baselines as follows.

BM25 BM25 is a sparse retrieval baseline (Robertson and Zaragoza, 2009).

DE w/ BM25 Neg Karpukhin et al. (2020) shows that the hard negatives from BM25 are more effective at training the dense retrievers than in-batch random negatives. With BM25's hard negatives, we train a dual-encoder as our first neural retriever.

CE w/ BM25 Neg We use BM25's hard negatives to train a cross-encoder as our first neural re-ranker.

CE w/ DE Neg CE w/ DE Neg is the second enhanced re-ranker. We follow Qu et al. (2021) to train CE w/ DE Neg. Specifically, we use CE w/ BM25 Neg to initialize the parameters, and use DE w/ BM25 Neg to retrieve negatives from the entire passage collection.

[§] https://fanyi.baidu.com

	MRR@10	Recall@1	Recall@50
BM25	21.03	12.08	70.00
DE w/ BM25 Neg	**53.96**	**41.53**	**91.33**

Table 4: Performance of retrieval models on the testing set of DuReader$_{\texttt{retrieval}}$.

BM25's top-50 psg.	MRR@10	Recall@1	Recall@50
CE w/ BM25 Neg	56.80	48.83	70.00
CE w/ DE Neg	57.62	51.52	70.00
DE's top-50 psg.	**MRR@10**	**Recall@1**	**Recall@50**
CE w/ DE Neg	**74.21**	**66.03**	**91.33**

Table 5: Performance of re-ranking models on testing set of DuReader$_{\texttt{retrieval}}$. We present re-ranking results based on two retrieval models including *BM25* and *DE w/ BM25 Neg*.

The relationships among our neural retrievers and re-rankers are shown in Figure 1. The training and architectural settings for all models are detailed in the Appendix A.2.

3.2 Evaluation Metrics

We use the following evaluation metrics in our experiments: (1) Mean Reciprocal Rank for the top 10 retrieved documents (MRR@10), (2) Recall for the top-1 retrieved items (Recall@1) and (3) Recall for the top-50 retrieved items (Recall@50). Recall@50 is more suitable for evaluating the first-stage retrievers, while MRR@10 and Recall@1 are more suitable for assessing the second-stage re-rankers.

3.3 Baseline Performance

We report the in-domain baseline performances for the first-stage retrievers in Table 4. Compared with the traditional retrieval system BM25, it is expected that DE w/ BM25 Neg outperforms the traditional system among all metrics, thanks to the powerful expressive ability of the neural encoder.

We then report the in-domain baseline performances for the second-stage re-rankers in Table 5. We observe that training the re-ranker with the hard negatives sampled from the neural retriever's top predictions is shown to outperform the negatives sampled from BM25's retrieved results in terms of MRR@10 and Recall@1.

3.4 Effects of Quality Improvements

In this section, we examine the effects of our strategies to improve the data quality of DuReader$_{\texttt{retrieval}}$ as in Section 2.3.

Model	MRR@10	Recall@1
	Duplicated	Duplicated
CE w/ Sim. Q	**50.6**	**43.93**
CE w/o Sim. Q	49.94	42.89

Table 6: Comparison of models by using two groups of training data: 1) **CE w/ Sim. Q**: training data without removing the queries that are semantically similar to the development and testing queries, 2) **CE w/o Sim. Q**: training data with removing the queries that are semantically similar to the development and testing queries. We evaluate the two models on the duplicated queries (**Duplicated**). All top-50 retrieval results are based on BM25. We **bold** the best model on each column.

Reducing False Negatives We test three models, including BM25, a dense retrieval model (DE w/ BM25 Neg) and a re-ranking model (CE w/ BM25 Neg) based on BM25's top-50 retrieved results, to quantify the impact of our strategy on reducing false negatives. Specifically, we compare the performance of the same model on the development set either with or without reducing false negatives. As shown in Figure 2, all metrics of the three models are significantly improved after adopting our strategy. These results suggest that there are many false negatives in the raw retrieval dataset derived from DuReader, and that our strategy successfully captures and reduces false negatives in development and testing sets.

Removing Similar Queries We conduct an experiment to quantify the effects of removing the training queries that are semantically similar to the development and testing queries. We train our re-ranking model (CE w/ BM25 Neg) by using the training data without (**CE w/o Sim. Q**) and with (**CE w/ sim. Q**) semantically duplicated queries, respectively. We then test both models on all 387 semantically duplicated queries (**Duplicated**) in the development and testing sets, as well as the rest of the development set (**Others**). We use BM25's top-50 retrieved results for the re-ranking models to re-rank. As shown in Table 6, comparing the two models' performance on Duplicated, we find model trained with those semantically similar queries (CE w/ Sim. Q) has a higher score on both MRR@10 and Recall@1. This suggests that using semantically similar queries in training may allow the model to simply memorize the data during training and achieve better performance during testing.

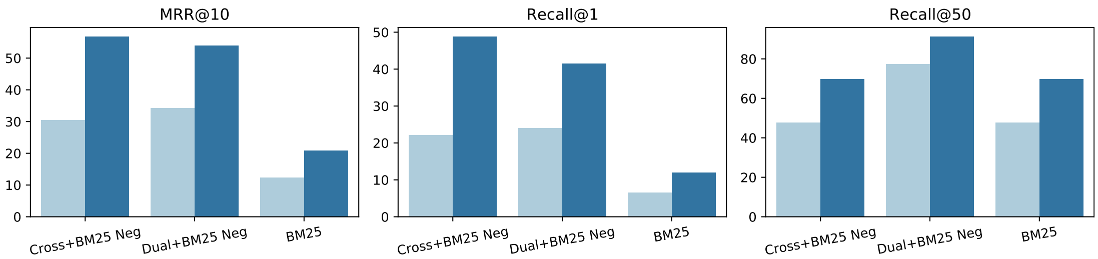

Figure 2: Comparison in the model performances before (light blue) and after (dark blue) reducing the false negatives in the development set via our pooling strategy and human annotation.

cMedQA	MRR@10		Recall@1		Recall@50	
	ZS.	FT.	ZS.	FT.	ZS.	FT.
BM25	6.26	\	4.98	\	14.05	\
DE	4.39	15.22	2.93	11.28	16.4	40.36

Table 7: Comparison for BM25 and the dense retrieval model DE w/ BM25 Neg (DE) in the Zero-shot (ZS.) experiments, and fine-tuning (FT.) experiments for estimating the upper-bound performance on cMedQA out-of-domain testing set.

cCOVID-News	MRR@10		Recall@1		Recall@50	
	ZS.	FT.	ZS.	FT.	ZS.	FT.
BM25	57.49	\	48.37	\	85.67	\
DE	14.02	56.67	9.91	46.68	38.04	87.78

Table 8: Comparison for BM25 and the dense retrieval model DE w/ BM25 Neg (DE) in the Zero-shot (ZS.) experiments, and fine-tuning (FT.) experiments for estimating the upper-bound performance on cCOVID-News out-of-domain testing set.

3.5 The Challenges and Limitations

In this section, we analyze the results of our best baseline system (i.e., retrieving the top-50 passages by DE w/ BM25 Neg, then re-ranking by CE w/ DE Neg) to better understand the specific challenges and limitations of $\text{DuReader}_{\text{retrieval}}$. Specifically, we manually analyze 500 query-passage predictions of the baseline. The 500 query-passage pairs are from 100 random-selected development queries with the top-5 passages retrieved and re-ranked by the baseline. To help understand the challenges and limitations of $\text{DuReader}_{\text{retrieval}}$, we ensure that the top-5 passages of these 100 queries contain no positive passages.

Salient Phrase Mismatch We observe that the mismatch in salient phrases between the query and the retrieved passages is particularly challenging for the baseline system as found in (Chen et al., 2021), accounting for 53.4% of total incorrect predictions. We further divide *salient phrase* into several sub-categories, i.e., *entity*, *numeral*, and *modifier*. Examples and explanations are in Table 10 in Appendix A.4.

Syntactic Mismatch We also observe that around 1% predictions have a syntactic mismatch between the query and the passage. The case in Table 10 in Appendix A.4 suggests that it is difficult for the baseline system to ensure the consistency in syntactic relationship between the query and the passages.

Other Challenges We also show two other typical challenges accounting for 22.6% incorrect predictions: 1) *Over-sensitivity on term overlap*: whether the baseline system is over-sensitive to retrieve the negative passages that contains a few lexical overlap with queries. 2) *robustness on typo*: whether the baseline system is robust against typos in queries or passages. Note that our dataset is constructed from the real query log of a commercial search engine. The noise in data (e.g. typos) challenges the robustness of the baseline system.

Limitations in False Negatives We notice that there are still about 14.8% false negatives. This suggests that despite the success of our strategy in Section 2.3 to reduce false negatives in development and testing sets to some extents, the presence of false negatives remains a challenge in building a high-quality passage retrieval benchmark.

3.6 Out-of-Domain Evaluation

We evaluate the out-of-domain (OOD) generalization ability of our dense retriever (DE w/ BM25 Neg) on the two OOD testing sets. We report the results in two settings: 1) Zero-shot setting: we directly evaluate DE w/ BM25 Neg without fine-tuning. 2) Fine-tuning setting: we fine-tune DE w/ BM25 Neg with the data from the target domain and evaluate it on OOD testing sets. The performance of the fine-tuned models is the estimated upper-bound that DE w/ BM25 Neg can achieve on

Model \ Evaluation	Monolingual	Cross-lingual
Supervised Model	-	28.03
Zero-shot Model	87.88	19.50
Transferred Model	-	**38.35**

Table 9: **Monolingual** (retrieving Chinese passages with Chinese queries) and **Cross-lingual** (retrieving Chinese passages with English queries) performance of the dual-encoder retrievers on our cross-lingual evaluations. We report the Recall@50 score for each retrieval model.

OOD testing sets.

In Table 7 and 8, we summarize the results of the out-of-domain experiments. First, we notice that the performance of the dense retriever is largely degraded on the two OOD testing sets. According to the in-domain evaluation (see Table 4 and 5), the dense retriever considerably outperforms BM25, however it has no obvious advantage over BM25 in the zero-shot setting, or even worse. In addition, the dense retriever can be significantly improved by fine-tuning. Its can maintain a large advantage over BM25 after fine-tuning on the target-domain. This results show that the dense retriever has limited domain transfer capability as observed in (Thakur et al., 2021).

3.7 Cross-lingual Evaluation

In the cross-lingual evaluation, we experiment with three dense retrieval models based on multilingual BERT (mBERT) (Devlin et al., 2019).

- **Supervised Model** We directly fine-tune mBERT using the parallel data of English queries and Chinese passages.

- **Zero-shot Model** We fine-tune an mBERT retriever on the full monolingual Chinese training data (i.e., 86K Chinese queries with Chinese positive paragraphs in DuReader$_{\mathrm{retrieval}}$), and directly evaluate it on the cross-lingual testing set.

- **Transferred Model** We further fine-tune **Zero-shot Model** by using the parallel data paired with English queries and Chinese passages, and then evaluate it on the cross-lingual testing set.

As shown in Table 9, we note that the performance of Zero-shot Model on cross-lingual testing set is less effective than Supervised Model. Furthermore, Zero-shot Model performs significantly worse on cross-lingual data than on monolingual data. According to these findings, cross-lingual retrieval is more difficult than monolingual retrieval, since the retriever cannot find relevant passages by simply matching shared terms between queries and passages (Litschko et al., 2021). Instead, cross-lingual retrievers must capture the semantic relevance of the query and passages. Additionally, Transferred Model outperformed other baselines, demonstrating the validity of transferring knowledge from the monolingual Chinese annotated data.

4 Related Works

Passage Retrieval Benchmarks. Passage retrieval and open-domain question-answering are challenging tasks that attracts much attention in developing the benchmarks. MS-MARCO (Nguyen et al., 2016) contains queries extracted from the search log of Microsoft Bing, which poses challenges in both the retrieval of relevant contexts and reading comprehension based on the contexts. Natural Questions (Kwiatkowski et al., 2019) is an open-domain question answering benchmark that consist of real queries issued to the Google search engine. These datasets are widely used for the research of passage retrieval. However, Lewis et al. (2021) find that there are 30% of test-set queries have semantically overlaps in the training queries for Natural Questions. Arabzadeh et al. (2021) observe that false negatives are common in MS-MARCO. TianGong-PDR (Wu et al., 2019) and Sougou-QCL (Zheng et al., 2018) are two Chinese retrieval datasets for the news documents and web-pages, separately. However, these datasets are either small or have no human annotation. Despite the progress in developing benchmarks for English passage retrieval, the large-scale and high-quality benchmarks for the non-English community are still limited.

Dense Retrieval Model. Information retrieval is a long-standing problem. In contrast to the traditional sparse retrieval methods (Salton and Buckley, 1988; Robertson and Zaragoza, 2009), recent dense retrievers aim at encoding the query and retrieved documents as contextualized representations based on the pre-training language models (Devlin et al., 2019; Sun et al., 2019), then calculate the relevance based on similarity function (Karpukhin et al., 2020; Luan et al., 2021; Qu et al., 2021) (e.g. cosine or dot product). Based on different learning paradigms, neural retrieval systems can be divided into two categories: 1) *unsupervised*: pre-training the retrieval without annotated data (Chang et al., 2020; Gao and Callan, 2021); 2) *supervised*: training the query and document encoders by contrasting the positives with designed negatives (Karpukhin et al., 2020; Xiong et al., 2021; Zhan et al., 2021). In terms of system architecture, the recent systems typically follow the two-stage framework (retrieval-then-re-ranking), in which a retriever (Mao et al., 2021; Nogueira et al., 2019; Dai and Callan, 2019) first retrieve a list of top candidates and the re-ranker (Gao et al., 2020; Khattab and Zaharia, 2020) will re-rank retrieved candidates. It has been shown that large-scale annotated datasets are one of the keys to successfully train dense retrievers (Karpukhin et al., 2020).

5 Conclusion

This paper presents a large-scale Chinese passage retrieval dataset to benchmark the retrieval systems. In order to ensure the quality of our dataset, we employ two strategies: 1) reducing the false negatives in development and testing sets using a pooling approach and human annotations, and 2) removing the training queries that are semantically similar to the development and testing queries. In addition, we provide two testing sets for out-of-domain evaluation, and a set for cross-lingual evaluation. We examine several retrieval baselines, including the traditional sparse retrieval system and the neural retrievers, and present the challenges and the limitations of our dataset. We hope this dataset can help facilitate the research of passage retrieval.

6 Limitations

As we discussed in Section 3.5, we still observe that approximately 14.8% of our best re-ranking model's wrong predictions are indeed caused by false negatives, even though we observed that our quality improvement strategy discussed in Section 2.3 was effective. This is primarily due to the difficulty of annotating the training data in a way that captures all positives.

Secondly, two out-of-domain testing sets are restricted to the medical domain. cMedQA focuses on the medical question-answering conversations, and cCOVID-NEWS focuses on the medical news domain. It may limit the ability to evaluate retrieval systems in other domains (e.g., the financial or legal domains).

7 Ethical Consideration

Our DuReader$_{\texttt{retrieval}}$ is developed only for research purpose. All data is collected from either the open-source projects, respecting corresponding licences' restrictions, or publicly available benchmarks. We do not guarantee that we have the copyright of this data, any may further discard data resources without copyright if necessary.

8 Acknowledgements

We thank the anonymous reviewers for their helpful comments. We also appreciate the feedback received from other members of the Deep Question Answering Team at Baidu NLP, specifically Qifei WU, for his valuable discussions and comments. This work is supported by the National Key Research and Development Project of China (No.2018AAA0101900).

References

Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles LA Clarke. 2021. Shallow pooling for sparse labels. *ArXiv preprint*, abs/2109.00062.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval. In *Advances in Neural Information Processing Systems*, volume 34, pages 7547–7560. Curran Associates, Inc.

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, Seattle, Washington, USA. Association for Computational Linguistics.

Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, , Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of ms marco passage ranking dataset.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *8th*

International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Xilun Chen, Kushal Lakhotia, Barlas Oguz, Anchit Gupta, Patrick S. H. Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2021. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? *ArXiv preprint*, abs/2110.06918.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 985–988. ACM.

Van Dang, Michael Bendersky, and W. Bruce Croft. 2013. Two-stage learning to rank for information retrieval. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 423–434, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *ArXiv preprint*, abs/2108.05540.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2020. Modularized transfomer-based ranking framework. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4180–4190, Online. Association for Computational Linguistics.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

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*, 7:452–466.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. On cross-lingual retrieval with multilingual text encoders. *ArXiv preprint*, abs/2112.11031.

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. *ArXiv preprint*, abs/1907.11692.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. 2019. Paddlepaddle: An open-source deep learning platform from industrial practice. *Frontiers of Data and Domputing*, 1(1):105–115.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.

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.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *ArXiv preprint*, abs/1904.08375.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

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

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-lingual training of dense retrievers for document retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 251–253, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *ArXiv preprint*, abs/1904.09223.

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)*.

Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*, volume 63. Citeseer.

Zhijing Wu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating passage-level relevance and its role in document-level relevance judgment. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 605–614. ACM.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. *Optimizing Dense Retrieval Model Training with Hard Negatives*, pages 1503–1512. Association for Computing Machinery, New York, NY, USA.

Jingtao Zhan, Xiaohui Xie, Jiaxin Mao, Yiqun Liu, M. Zhang, and Shaoping Ma. 2022. Evaluating extrapolation performance of dense retrieval. *ArXiv*, abs/2204.11447.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Yukun Zheng, Zhen Fan, Yiqun Liu, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Sogou-qcl: A new dataset with click relevance label. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1117–1120. ACM.

Hongyu Zhu, Yan Chen, Jing Yan, Jing Liu, Yu Hong, Ying Chen, Hua Wu, and Haifeng Wang. 2021. Duqm: A chinese dataset of linguistically perturbed natural questions for evaluating the robustness of question matching models. *ArXiv preprint*, abs/2112.08609.

A Appendix

A.1 Details for Re-ranker Used in Reducing False Negatives

We first use four different pre-training models, including ERNIE (Sun et al., 2019), BERT (Devlin et al., 2019; Cui et al., 2021), RoBERTa (Liu et al., 2019) and MacBERT (Cui et al., 2020), as the initializations to train four cross-encoder re-rankers as in (Qu et al., 2021) with negatives sampled from the pooled passages as discussed in Section 2.3. We then ensemble these four re-ranking models by averaging their prediction scores for each query-passage pair.

A.2 Baseline Implementation Details

We conduct all experiments with the deep learning framework PaddlePaddle (Ma et al., 2019) on up to eight NVIDIA Tesla A100 GPUs (with 40G RAM).

We use the ERNIE 1.0 base (Sun et al., 2019) as the initializations for both our first dual-encoder retriever (DE w/ BM25 Neg) and cross-encoder re-ranker (CE w/ BM25 Neg). ERNIE shares the same architecture with BERT but is trained with entity-level masking and phrase-level masking to obtain better knowledge-enhanced representations. To train our second enhanced re-ranker (CE w/ DE Neg), we use the parameters from CE w/ BM25 Neg as initialization.

For training settings, we also use the Cross-batch negatives setting as in (Qu et al., 2021). When sampling the hard negatives from the top-50 retrieved items, we sample 4 negatives per positive passage. The dual-encoders are trained with the batch size of 256. The cross-encoders are trained with the batch size of 64. The dual-encoders and cross-encoders are trained with 10 and 3 epochs. We use ADAM optimizer for all models' trainings and the learning rate of the dual-encoder is set to 3e-5 with the rate of linear scheduling warm-up at 0.1, while the learning rate of the cross-encoder is set to 1e-5 with no warm-up training. We set the maximal length of questions and passages as 32 and 384, respectively.

In inference time of our dense retrieval model (DE w/ BM25 Neg), we use FAISS (Johnson et al., 2019) to index the dense representations of all passages.

A.3 Details for Human Annotations

We perform the annotation in our internal annotation platform to ensure the data quality, where all the annotators and reviewers are full-time employees. The pairs of all queries and their pooled top-5 paragraphs retrieved by all models are divided into packages, with 1K samples for each. Annotators are asked to identify whether each query-paragraph pair is relevant for a single package. Then at least two reviewers check the accuracy of this package by reviewing 100 random query-paragraph pairs independently. If the average accuracy is less than the threshold (i.e., 93%), the annotators will be asked to revise the package until the accuracy is higher than the threshold.

A.4 Cases for Challenges in Error Analysis

We present the selected cases in Table 10 and discuss them in this section to support our error analysis in Section 3.5.

Salient Phrase Mismatch Taking the entity mismatch as an example, we expect that the main entity in the retrieved passage should be consistent with the query. However, the second example in Table 10 shows that the query asks for information about *Taobao*, but the retrieved passage is related to *Alipay* instead. There is a challenge for retrieval systems to filter out passages that entail entities inconsistent with the query.

Syntax Mismatch Given the case showed in Table 10 as an example, the retrieval system is hard to understand the subject and object in the example query are *Taipei* and *Ruifang*, instead, it simply ranks the candidate passage entailing *Taipei* and *Ruifang* to a top predictions.

Other Challenges In our analysis, it is found that about 21% of the errors are due to the retrieval system simply predicting its output based on the presence of co-occurring low-frequency terms (e.g., "*wow*" in the example in Table 10) in query and paragraph, but their semantic meanings are not related indeed. And about 1.6% of the errors are due to noise in the query or paragraph. For example, misspelling the "*iPhone*" as "*ipone*".

Category	Type	Example Query	Example Passage	Explanation	%
Salient Phrase Mismatch	Entity mismatch	淘宝修改实名认证 Change authentication name at Taobao	响应国家规定，支付宝即将实施支付宝个人信息实名认证... In response to national regulations, Alipay will soon implement real-name authentication...	The entities in the query (Taobao) and the passage (Alipay) are mismatched.	39%
	Numeral mismatch	最近有什么好听的歌2016 Any nice songs in 2016	...音乐巴士2017好听的歌榜单收藏了你最需要的的好歌... ...Music Bus's Best Songs 2017 has the songs you need most...	The query asks for songs in 2016, but the passage is about songs in 2017.	5%
	Modifier mismatch	吃完海鲜可以喝牛奶吗 Can I drink milk after having seafood?	不可以。早晨喝牛奶，不科学。原因是… No. Drinking milk in the morning is not good for health. The reason is...	The modifier in query (after eating seafood) and the one in the passage (in the morning) are different.	9.4%
Syntactic mismatch	Syntactic mismatch	台北怎么去瑞芳 How to go from Taipei to Ruifang	从瑞芳回台北，乘火车1小时到达，可以使用悠游卡... It takes 1 hour by train from Ruifang to Taipei. You can use the EasyCard...	The query asks how to go from Taipei to Ruifang but the paragraph is about going from Ruifang to Taipei.	1%
Other Challenges	Robustness on typos	~~ipone~~iPhone手机屏幕右上角有个圈是什么 What is the circle in the upper right corner of the ~~ipone~~iPhone screen	但是又不知道这些图标是怎么出现的。是什么东西，干什么用的，例如手机信号格旁边的眼睛图标是代表了开启只能屏幕… But I don't know how these icons appeared. What is it and what is it for? For example, the eye icon next to the mobile phone signal grid means that the screen can only be turned on...	Typos may introduce noise to the model's understanding of query or passage, e.g., it may affects the identification of the main entity iPhone in the query.	1.6%
	Over-sensitivity on term overlap	wow邮件发错了怎么办 What should I do if my wow email is sent by mistake	...还有个方法就是你登陆WOW然后都设置好后退出游戏… ...and another way is that you log in to WOW and then exit the game after setting everything up...	The query and passage (i.e., WOW) has a matched term in ; but they are semantically irrelevant.	21%
Others	False negatives	求推荐好看的国产电视剧 Please recommend good Chinese TV series	TOP3：美人制造。30集全。主演：杨蓉金世佳邓萃雯。简介：以唐代女皇武则天时期为大背景... TOP3: Beauty Made. 30 episodes. Starring: Yang Rong, Jin Shijia, Deng Cuiwen. Introduction: Taking the period of Empress Wu Zetian of the Tang Dynasty as the background...	\	14.8%
	Others	\	\	Cases not fitting into any of above categories.	8.2%

Table 10: Summary of the manual analysis for the 500 query-passage pairs predicted by our strongest re-ranker (CE w/ DE Neg). We highlight the challenges in *salient phrase mismatch* in red, *syntax mismatch* in blue, and *Other Challenges* in green.