Hyperlink-induced Pre-training for Passage Retrieval of Open-domain Question Answering

Anonymous ACL submission

Abstract

To alleviate the data scarcity problem in training question answering systems, recent works propose additional intermediate pre-training for dense passage retrieval (DPR). However, there still remains a large discrepancy between the provided upstream signals and the downstream question-passage relevance, which leads to less improvement. To bridge this gap, we propose the **H**yper**L**ink-induced **P**re-training (HLP), a method to pre-train the dense retriever with the text relevance induced by hyperlink-based topology within Web documents. We demonstrate that the hyperlink-based structures of *dual-link* and *co-mention* can provide effective relevance signals for large-scale pre-training that better facilitate downstream passage retrieval. We investigate the effectiveness of our approach across a wide range of open-domain QA datasets under zero-shot, few-shot, multi-hop, and out-of-domain scenarios. The experiments show our HLP outperforms the BM25 by up to 7 points as well as other pre-training methods by up to 30 points in terms of top-20 retrieval accuracy under the zero-shot scenario. Furthermore, HLP significantly outperforms other pre-training methods under the other scenarios.

1 Introduction

Open-domain question answering (OpenQA) aims to answer factual open questions with a large external corpus of passages. Current approaches to OpenQA usually adopt a two-stage retriever-reader paradigm (Chen et al., 2017; Zhu et al., 2021) to fetch the final answer span. The performance of OpenQA systems is largely bounded by the retriever as it determines the evidential documents for the reader to examine. Traditional retrievers, such as TF-IDF and BM25 (Robertson and Zaragoza, 2009), are considered incapable of adapting to scenarios where deep semantic understanding is required. Recent works (Lee et al., 2019; Karpukhin et al., 2020; Qu et al., 2021) show that by fine-tuning pre-trained language models on sufficient downstream data, dense retrievers can significantly outperform traditional term-based retrievers.

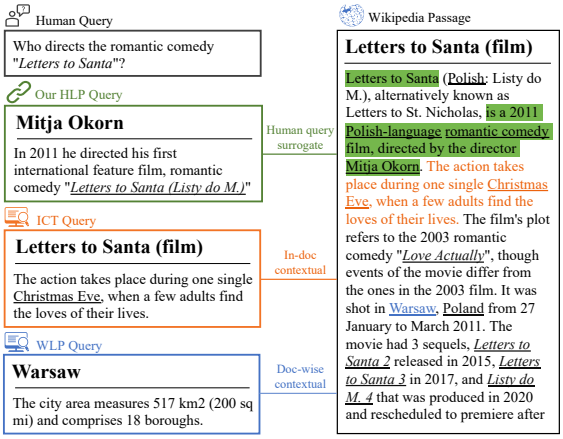

Figure 1: An example of different kinds of pseudo Q-P pairs. Underlined texts are hypertexts that linked to other Wikipedia pages. The ICT query is a random sentence originated from the passage and the WLP query is a sentence from the first section of an out-link document of the given passage. The text highlighted in green gives evidence to answer the human query, and our proposed HLP query can be a better surrogate of the human query.

Considering the data-hungry nature of the neural retrieval models, extensive efforts (Lee et al., 2019; Chang et al., 2020; Sachan et al., 2021) have been made to design self-supervised tasks to pre-train the retriever. However, these pre-training tasks construct relevance signals largely depending on easily achieving sentence-level or document-level contextual relationships. For example, the relationship between a sentence and its originated context (shown by the ICT query in Figure 1) may not be sufficient enough to facilitate question-passage matching for the tasks of OpenQA. We also find that these pre-trained retrievers still fall far behind BM25 in our

We will soon release our source code and pre-training corpus on github.

pilot study on the zero-shot experiment.

In order to address the shortcomings of the matching-oriented pre-training tasks as mentioned above, we propose a pre-training method with better surrogates of real natural question-passage (Q-P) pairs. We consider two conditions of relevance within Q-P pairs, which is similar to the process of distantly supervised retriever learning (Mintz et al., 2009; Chen et al., 2017).

1) **Evidence Existence** The evidence, such as entities and their corresponding relations, should exist across the query and the targeted passage as they both discuss similar facts or events related to the answer.
2) **Answer Containing** The golden passage should contain the answer of the query, which means that a text span within the passage can provide the information-seeking target of the query.

In this paper, we propose **H**yper**L**ink-induced **P**re-training (HLP), a pre-training method to learn effective Q-P relevance induced by the hyperlink topology within naturally-occurring Web documents. Specifically, these Q-P pairs are automatically extracted from the online documents with relevance adequately designed via hyperlink-based topology to facilitate downstream retrieval for question answering. Figure 1 shows an example of comparison between the human-written query and different pseudo queries. By the guidance of hyperlinks, our HLP query hold the relevance of answer containing with the passage (query title occurs in the passage). Meanwhile, the HLP query can introduce far more effective relevance of evidence existence than other pseudo queries by deeply mining the hyperlink topology, e.g., the dual-link structure. In figure 1, both HLP query and the passage both contain information corresponding to the same fact of *"Mitja Okorn directed the film of Letters to Santa"*. This makes our pseudo query low-cost and a good surrogate for the manually written query.

Our contributions are two-fold. First, we present a hyperlink-induced relevance construction methodology that can better facilitate downstream passage retrieval for question answering, and specifically, we propose a pre-training method: Hyperlink-induced Pre-training (HLP). Second, we conduct evaluations on six popular QA datasets, investigating the effectiveness of our approach under zero-shot, few-shot, multi-hop, and out-of-domain (OOD) scenarios. The experiments show HLP outperforms BM25 in most of the cases under the zero-shot scenario and other pre-training methods under all scenarios.

2 Related Work

Dense Retriever Pre-training Previous works have attempted to conduct additional pre-training for dense retrievers on various weakly supervised data. Borisov et al. (2016) and Dehghani et al. (2017) pre-trained ranking models on click-logs and BM25-induced signals respectively for web search. Lee et al. (2019) proposed the inverse cloze task (ICT) to pre-train a dense retrieval model, which randomly selects sentences as pseudo queries, and matched them to the passages that they originate from. Besides, Chang et al. (2020) proposed the pre-training task of wiki link prediction (WLP) and body first selection (BFS) tasks. Similar to our work, the WLP task also leveraged the hyperlinks within Wikipedia to construct relevant text pairs. However, as shown in figure 1, the WLP pseudo query can only ensure the weak doc-wise contextual relationship with the passage. Guu et al. (2020) proposed the masked-salient-span pre-training task which optimizes a retrieval model by the distant supervision of language model objective. As a follow-up, Sachan et al. (2021) combined ICT with the masked-salient-span task and further improved the pre-training effectiveness.

Data Augmentation via Question Generation Ma et al. (2021), Reddy et al. (2021) and Oğuz et al. (2021) all investigate training a dense retriever on questions synthesized by large question generative (QG) models. Targeting on the zero-shot setting, Ma et al. (2021) trained a question generator on general-domain question passage pairs from community platforms and publicly available academic datasets. Reddy et al. (2021) focused more on domain transfer and trained the QG model on QA datasets of Wikipedia articles. Oğuz et al. (2021) uses the synthetically generated questions from PAQ datasaset (Lewis et al., 2021) and the post-comment pairs from dataset of Reddit conversations for retrieval pre-training. Recently, Shinoda et al. (2021) reveals that the QG models tend to generate questions with high lexical overlap which amplify the bias of QA dataset. Different to these studies, our method focuses on a more general setting where the retriever is only trained with the naturally occurring web documents, and has no

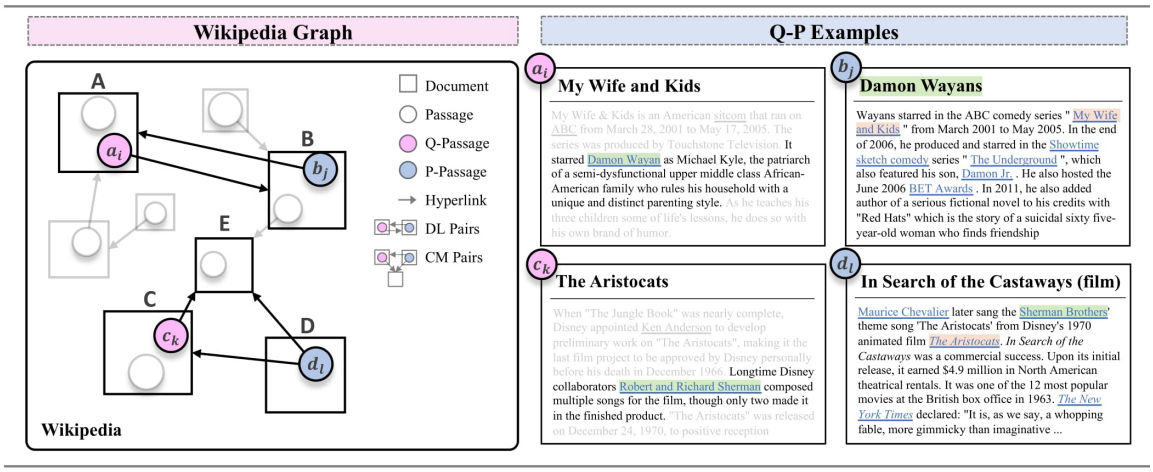

Figure 2: The figure on the left shows a partial Wikipedia graph where two types of pseudo Q-P pairs (a_i, b_j) and (c_k, d_l) are presented. The text boxes on the right show two concrete examples of HLP Q-P pairs where text highlighted in green gives evidence while in orange indicates the answer span.

access to any downstream datasets.

3 Hyperlink-induced Pre-training (HLP)

In this section, we firstly discuss the background of OpenQA retrieval, then our methodology and training framework.

3.1 Preliminaries

Passage Retrieval Given a question q, passage retrieval aims to provide a set of relevant passages p from a large corpus $\mathcal{D}$. Our work adopts Wikipedia as source corpus and each passage is a disjoint segment within a document from $\mathcal{D}$.

OpenQA Q-P Relevance For OpenQA, a passage p is considered relevant to the query q if p conveys similar facts and contains the answer to q. These two conditions of relevance, namely evidence existence and answer containing, are properly introduced into the HLP Q-P pairs under the guidance of desired hyperlink structure. We will discuss more in this section.

To better formulate the relevance of pseudo Q-P pairs, we denote the sequence of passages within a document as $A = [a_1, a_2, ..., a_{n_A}]$ where $A \in \mathcal{D}$. The corresponding topical entity and the title of document A are denoted as e_A and t_A, respectively. We use m_A to indicate a mention of entity e_A, which is a hypertext span linking to document A. Note that the mention span m_A is usually identical to the document title t_A or a variant version of it. Further, we define $\mathcal{F}_{(p)}$ as the entity-level factual information conveyed by the passage p, which is a set consists of the topical entity e_P and the entities mentioned within passage p.

Evidence Existence in HLP With appropriately designed hyperlink topologies, our HLP Q-P pairs guarantee the co-occurrence of entities which are presented as hypertext or topics in q and p. This is considered as evidence across the Q-P pairs:

$$\mathcal{F}_{(q)} \cap \mathcal{F}_{(p)} \neq \emptyset \tag{1}$$

Furthermore, we conjecture that HLP is more likely to achieve fact-level relevance than entity-level overlap. We conduct human evaluation in Section 6.3 and case studies in Appendix G to support this conjecture. Moreover, we demonstrate that any Q-P pair containing hyperlink-induced factual evidence, which is represented as a triple that induced by hyperlinks, is included in our proposed topologies, which are included in Appendix D.

Answer Containing in HLP We consider the document title t_Q as the information-seeking target of q. Accordingly, the relevance of answer containing can be formulated as

$$t_Q \subseteq p \tag{2}$$

The rationale behind this is that both the natural question and the Wikipedia document are intended to describe related facts and events regarding a targeted object, whereas the object is an answer for a question but a topic for a Wikipedia document. This similarity leads us to take the document title as the information-seeking target of its context.

3.2 Hyperlink-induced Q-P Pairs

Based on analysis of how queries match their evidential passages in the NQ (Kwiatkowski et al.,

2019) dataset, we propose two kinds of hyperlink topology for relevance construction: Dual-link and Co-mention. We present our exploratory data analysis on NQ dataset in Appendix C. Here we discuss the desired hyperlink topologies and the corresponding relevance of the pseudo Q-P pairs.

Dual-link (DL) Among all NQ training samples, 55% of questions mention the title of their corresponding golden passage. This observation motivates us to leverage the topology of dual-link (DL) for relevance construction. We consider a passage pair (a_i, b_j) follows the dual-link topology if they link to each other. An example of a DL pair (a_i, b_j) is shown in Figure 2, in which passage b_j mentions the title of document A as m_A, satisfying the condition of answer containing:

$$t_A \approx m_A \quad \text{and} \quad m_A \subseteq b_j \tag{3}$$

Further, since the passages a_i and b_j both mention the topical entity of the other, the entities e_A and e_B appear in both passages as evidence:

$$\{e_A, e_B\} \subseteq \mathcal{F}_{(a_i)} \cap \mathcal{F}_{(b_j)} \tag{4}$$

Co-mention (CM) Among all NQ training samples, about 40% of questions fail to match the dual-link condition but mention the same third-party entity as their corresponding golden passages. In light of this observation, we utilize another topology of Co-mention (CM). We consider that a passage pair (c_k, d_l) follows the Co-mention topology if they both link to a third-party document E and d_l links to c_k. Figure 2 illustrates a CM pair (c_l, d_k) where answer containing is ensured as the title of c_k occurs in d_l:

$$t_C \approx m_C \quad \text{and} \quad m_C \subseteq d_l \tag{5}$$

Since both c_l and d_k mention a third-party entity e_E, and that e_C is a topical entity in c_l while a mentioned entity in d_k, we have entity-level evidence across c_l and d_k as:

$$\{e_C, e_E\} \subseteq \mathcal{F}_{(c_k)} \cap \mathcal{F}_{(d_l)} \tag{6}$$

In practice, we use sentence-level queries which contain the corresponding evidential hypertext, and we do not prepend the title to the passage in order to reduce the superficial entity-level overlap. To improve the quality of CM pairs, we filter out those with a co-mentioned entity which has a top 10% highest-ranked in-degree among the Wikipedia entity. To illustrate how we construct our pseudo Q-P pairs, we present pseudo code in Appendix E.

Furthermore, we highlight that HLP has the following advantages: 1) it introduces more semantic variants and paraphrasing for better text matching. 2) The hypertext reflects potential interests or needs of users in relevant information, which is consistent to the downstream information-seeking propose.

3.3 Bi-encoder Training

We adopt a BERT-based bi-encoder to encode queries and passages separately into d-dimension vectors. The output representation is derived from the last hidden state of the [CLS] token and the final matching score is measured by the inner product:

$$\begin{aligned} h_q &= \mathrm{BERT_Q}(q)([\mathrm{CLS}]) \\ h_p &= \mathrm{BERT_P}(p)([\mathrm{CLS}]) \\ \mathrm{S}(p, q) &= h_q^{\mathrm{T}} \cdot h_p \end{aligned}$$

Let $B = \{\langle q_i, p_i^+, p_i^- \rangle\}_{i=1}^n$ be a mini-batch with n instances. Each instance contains a question q_i paired with a positive passage p_i^+ and a negative passage p_i^-. With in-batch negative sampling, each question q_i considers all the passages in B except its own gold p_i^+ as negatives, resulting in $2n-1$ negatives per question in total. We use the negative log likelihood of the positive passage as our loss for optimization:

$$\begin{aligned} & L(q_i, p_i^+, p_{i,1}^-, ..., p_{i,2n-1}^-) \\ = & -log \frac{e^{S(q_i, p_i^+)}}{e^{S(q_i, p_i^+)} + \sum_{j=1}^{2n-1} e^{S(q_i, p_{i,j}^-)}} \end{aligned}$$

4 Experimental Setup

In this session, we discuss the pre-training corpus preparation, downstream datasets, the hyper-parameter and the basic setup for our experiments.

4.1 Pre-training Corpus

We adopt Wikipedia as our source corpus $\mathcal{D}$ for pre-training as it is the largest encyclopedia covering diverse topics with good content quality and linking structures. We choose the snapshot 03-01-2021 of an English Wikipedia dump, and process it with WikiExtractor[2] to obtain clean context. After filtering out documents with blank text or a title less than three letters, following previous work (Karpukhin et al., 2020), we split the remaining documents into disjoint chunks of 100 words as passages, resulting in over 22 million passages in the end.

[2] Available at https://github.com/attardi/wikiextractor

4.2 Downstream Datasets

We evaluate our method on several open-domain question answering benchmarks which are shown below.

Natural Questions (NQ) (Kwiatkowski et al., 2019) is a popular QA dataset with real queries from Google Search and annotated answers from Wikipedia.

TriviaQA (Joshi et al., 2017) contains question-answer pairs scraped from trivia websites.

WebQuestions (WQ) (Berant et al., 2013) consists of questions generated by Google Suggest API with entity-level answers from Freebase.

HotpotQA (Fullwiki) (Yang et al., 2018) is a human-annotated multi-hop question answering dataset.

BioASQ (Tsatsaronis et al., 2015) is a competition on biomedical semantic indexing and question answering. We evaluate its factoid questions from task 8B.

MS MARCO (Passage Ranking) (Nguyen et al., 2016) consists of real-world user queries and a large collection of Web passages extracted by Bing search engine.

Retrieval Corpus For downstream retrieval, we use the 21M Wikipedia passages provided by DPR (Karpukhin et al., 2020) for NQ, TriviaQA and WQ. For BioASQ, we take the abstracts of PubMed articles from task 8A with the same split to Reddy et al. (2021)'s work. For HotpotQA and MS MARCO, we use the official corpus.

4.3 Implementation Details

During the pre-training, we train the bi-encoder for 5 epochs with parameters shared, using a batch size of 400 and an Adam optimizer (Kingma and Ba, 2014) with a learning rate 2×10^{-5}, linear scheduling with 10% warm-up steps. Our HLP and all the reproduced baselines are trained on 20 million Q-P pairs with in-batch negative sampling, and the best checkpoints are selected based on the average rank of gold passages evaluated on the NQ dev set. The pre-training takes around 3 days using eight NVIDIA V100 32GB GPUs.

For the downstream, we use the same hyper-parameters for all experiments. Specifically, we fine-tune the pre-trained models for 40 epochs with a batch size of 256 and the same optimizer and learning rate settings to the pre-training. We conduct evaluation on respective dev sets to select best checkpoints, and we use the last checkpoint if there is no dev set or test set (e.g. HotpotQA). More details can be found in the Appendix A.

4.4 Baselines

Most existing baselines have been implemented under different experimental settings, which have a substantial effect on the retrieval performance. To ensure fairness, we reproduce several pre-training methods (ICT, WLP, BFS, and their combination) under the same experimental setting, such as batch size, base model, amount of pre-training data, and so on. The only difference between our method and the re-implemented baselines is the self-supervision signal derived from the respective pre-training samples. Our reproduced BM25 baseline is better than that reported in Karpukhin et al. (2020), and the re-implemented pre-training methods also perform better than those reported by the recent work[3]. In addition, we include the work REALM (Guu et al., 2020) as a baseline which has recently been reproduced by Sachan et al. (2021) using 240 GPUs and is named masked salient spans (MSS). We note that most related works gain improvements from varying downstream setting or synthetic pre-training with access to the downstream data of respective domain, which is out of the scope of our interests.

5 Experiments

5.1 Main Results

Table 1 shows the retrieval accuracy of different models on three popular QA datasets under zero-shot and full-set fine-tuning settings.

Under zero-shot setting, HLP consistently outperforms BM25 except for the top-5 retrieval accuracy of TriviaQA, while all other pre-training baselines are far behind. We attribute the minor improvement over BM25 on TriviaQA to a high overlap between questions and passages, which gives term-based retriever a clear advantage. We investigate the coverage of the question tokens that appear in the gold passage and find that the overlap is indeed higher in TriviaQA (62.8%) than NQ (60.7%) and WQ (57.5%).

After fine-tuning, all models with intermediate pre-training give better results than the vanilla DPR while our HLP achieves the best in nearly all cases.

[3]Our reproduced ICT and BFS surpass the reproduction from recent work (Oğuz et al., 2021) by 15 and 12 points, respectively, in terms of top-20 retrieval accuracy on NQ test set under zero-shot setting.

	NQ			TriviaQA			WQ		
	top5	top20	top100	top5	top20	top100	top5	top20	top100
	w/o fine-tuning (zero-shot)								
BM25†	43.6	62.9	78.1	**66.4**	76.4	83.2	42.6	62.8	76.8
ICT† (Lee et al., 2019)	23.4	40.7	58.1	33.3	51.3	69.9	19.9	36.2	56.0
WLP† (Chang et al., 2020)	28.5	47.3	65.3	51.3	67.0	79.1	26.9	49.0	68.1
BFS† (Chang et al., 2020)	31.0	49.9	67.5	43.8	61.1	74.7	28.5	48.0	67.7
ICT+WLP+BFS† (Chang et al., 2020)	32.3	50.2	68.0	49.7	65.5	78.3	28.4	47.8	67.5
MSS (Sachan et al., 2021)	41.7	59.8	74.9	53.3	68.2	79.4	-	-	-
HLP	**51.2**	**70.2**	**82.0**	65.9	**76.9**	**84.0**	**49.3**	**66.9**	**80.8**
	w/ fine-tuning								
No Pre-train†	68.5	79.6	86.5	71.3	79.7	85.0	61.6	74.5	81.7
ICT† (Lee et al., 2019)	69.8	81.1	87.0	70.4	79.8	85.5	63.7	75.5	83.4
WLP† (Chang et al., 2020)	69.8	**81.4**	87.4	73.1	81.5	86.1	64.5	75.2	83.9
BFS† (Chang et al., 2020)	68.7	80.1	86.5	72.8	80.8	86.0	63.0	75.1	83.5
ICT+WLP+BFS† (Chang et al., 2020)	68.9	80.9	87.7	74.6	82.2	86.5	64.1	**76.7**	84.4
HLP	**70.9**	**81.4**	**88.0**	**75.3**	**82.4**	**86.9**	**65.5**	76.5	**84.5**

Table 1: Top-k ($k \in \{5, 20, 100\}$) retrieval accuracy, measured as the percentage of top k retrieved passages with the answer contained. The upper block of the table describes the performance under zero-shot setting, while the lower under the full-set fine-tuning setting. †: Our re-implementation.

Among ICT, WLP and BFS, we observe that WLP is the most competitive with or without fine-tuning, and additional improvements can be achieved by combining three of them. This observation indicates that pre-training with diverse relevance leads to better generalization to downstream tasks, while document-wise relevance is more adaptable for the OpenQA retrieval. The advantage of document-wise relevance may come from the fact that texts in different documents are likely written by different parties, providing less superficial cues for text matching, which is beneficial for the downstream retrieval. Our HLP learns both coarse-grained document-wise relationships as well as the fine-grained entity-level evidence, which results in a significant improvement.

5.2 Few-shot Learning

To investigate the retrieval effectiveness in a more realistic scenario, we conduct experiments for few-shot learning. Specifically, we fine-tune the pre-trained models on large datasets (NQ, TriviaQA) with m ($m = \{16, 256, 1024\}$) samples and present the few-shot retrieval results in Table 2. With only a few hundred labeled data for fine-tuning, all the models with intermediate pre-training perform better than that without, and HLP outperforms the other methods by a larger margin when m is smaller. Among three re-implemented baselines, WLP gains the largest improvement with increasing number of samples, outperforming ICT and BFS when a thousand labelled samples are provided for fine-tuning.

	NQ			TriviaQA		
	top5	top20	top100	top5	top20	top100
	m = 16					
No Pre-train	12.7	24.2	40.2	18.6	32.6	51.0
ICT	37.1	54.4	70.5	47.2	62.5	75.8
WLP	29.8	48.2	65.5	51.4	66.9	79.2
BFS	39.8	57.9	73.2	46.9	62.2	75.2
HLP	**51.9**	**70.3**	**81.6**	**65.9**	**76.9**	**84.0**
	m = 128					
No Pre-train	38.0	53.4	68.8	38.0	53.4	68.8
ICT	47.0	64.2	77.4	58.5	71.4	81.0
WLP	44.9	62.4	76.6	63.1	74.5	82.6
BFS	44.4	62.8	76.7	59.2	71.7	80.8
HLP	**55.2**	**71.3**	**81.8**	**67.7**	**77.7**	**84.4**
	m = 1024					
No Pre-train	49.7	66.4	78.8	54.0	67.2	77.6
ICT	55.9	72.2	83.7	63.8	75.7	83.3
WLP	57.2	73.6	83.9	67.2	77.5	84.5
BFS	53.7	71.7	83.1	63.6	75.3	83.1
HLP	**60.6**	**76.4**	**85.3**	**70.2**	**79.8**	**85.4**

Table 2: Few-shot retrieval accuracy on NQ and TriviaQA test sets after fine-tuning with m annotated samples.

5.3 Out-of-domain (OOD) Scenario

While HLP is pre-trained on Wikipedia pages, we conduct additional experiments on BioASQ and MS MARCO datasets with non-Wikipedia corpus to further verify its out-of-domain (OOD) generalization. Following Gururangan et al. (2020), we measure the similarity between corpus by computing the vocabulary overlap of the top 10K frequent words (excluding stopwords). We observe a vocabulary overlap of 36.2% between BioASQ and Wikipedia while 61.4% between MS MARCO and Wikipedia, indicating that these two domains differ considerably from our pre-training corpus.

The results of zero-shot retrieval on BioASQ

Model	Negative	NQ			TriviaQA			WebQ		
		top5	top20	top100	top5	top20	top100	top5	top20	top100
Dual-link	0	46.2	64.7	78.0	60.5	73.0	81.2	44.6	65.2	78.8
	1	49.0	67.8	79.7	62.0	73.8	82.1	48.4	**67.1**	79.5
Co-mention	0	35.8	57.1	75.1	58.9	73.1	82.6	36.2	58.9	76.2
	1	42.5	62.2	77.9	63.2	75.8	83.7	45.4	64.5	78.9
HLP	0	45.7	66.0	79.9	62.6	75.2	83.0	43.9	64.1	79.4
	1	**51.2**	**70.2**	**82.0**	**65.9**	**76.9**	**84.0**	**49.3**	66.9	**80.8**

Table 3: Ablation studies on different types of topologies and negatives. The retrieval accuracy of models trained with different types of Q-P pairs and additional negatives on NQ, TriviaQA, and WebQ datasets.

and MS MARCO datasets are presented in Table 4. For BioASQ, HLP is competitive with both BM25 and AugDPR(Reddy et al., 2021) while significantly outperforming ICT, WLP, and BFS. Note that AugDPR is a baseline that has access to NQ labeled data whereas our HLP is trained in an unsupervised way. For MS MARCO, HLP consistently outperforms other pre-training methods but falls behind BM25 under zero-shot setting. We conjecture the performance degradation on MS MARCO is attributed to two factors: 1) the Q-P lexical overlap of MS MARCO (65.7%) is higher than that in BioASQ (48.7%) as well as other datasets; 2) the information-seeking target of the MS MARCO query is the entire passage rather than a short answer span, which is biased towards our proposed answer containing. we also observe that pre-training exclusively with DL pairs achieves better results in MS MARCO, indicating the generality of relevance induced by DL topology.

	BioASQ		MS MARCO	
	top20	top100	R@20	R@100
BM25	42.1‡	50.5‡	**49.0**	**69.0**
DPR	34.7‡	46.9‡	-	-
AugDPR	41.4‡	52.4‡	-	-
ICT	8.9	18.6	10.8	19.5
WLP	29.7	44.3	18.4	36.0
BFS	28.4	41.9	28.0	44.7
HLP (DL)	**46.0**	56.9	42.0	62.6
HLP (CM)	37.8	54.7	26.6	47.3
HLP (DL+CM)	40.8	**58.3**	37.3	60.0

Table 4: Top-20/100 zero-shot retrieval accuracy on BioASQ and Top-20/100 zero-shot recall on MS MARCO. ‡: (Reddy et al., 2021)

5.4 Multi-hop Retrieval

While HLP aims to acquires the ability in matching document-wise concepts and facts, it raises our interest in its capability for multi-hop scenarios. We evaluate our methods on HotpotQA in a single-hop manner. Specifically, for each query, we randomly selects one golden passage from the two as a positive passage and one additional passage with high TF-IDF scores as a negative passage. Our models are further fine-tuned on the HotpotQA training set and evaluated on the bridge and the comparison type questions from the development set, respectively. The results of our study are shown in Table 5 which reveals that HLP consistently outperforms others methods, with up to a 11-point improvement on top-5 retrieval accuracy of bridge questions. Furthermore, WLP yields a 4-point advantages in average over ICT and BFS on bridge questions, showing that document-wise relevance contributes to better associative abilities. We include a case study in Appendix F.

	Bridge			Comparison		
	top5	top20	top100	top5	top20	top100
No Pre-train	25.0	40.5	58.0	83.0	94.2	97.4
ICT	28.1	43.8	61.8	84.8	94.4	98.3
WLP	32.1	49.1	66.0	89.7	97.3	99.2
BFS	29.0	44.7	62.1	87.4	95.8	98.7
HLP	**36.9**	**53.0**	**68.5**	**94.4**	**98.5**	**99.5**

Table 5: Retrieval accuracy on questions from HotpotQA dev set, measured as the percentage of top-k retrieved passages which include both golds.

6 Analysis

6.1 Ablation Study

To better understand how different key factors affect the results, we conduct ablation experiments with results shown in Table 3.

Hyperlink-based Topologies Our proposed dual-link (DL) and co-mention (CM) Q-P pairs, provide evidence induced by different hyperlink-based topologies. To examine their respective effectiveness, we pre-train retrievers on Q-P pairs derived from each topology and their combinations. We present zero-shot retrieval results in Table 3, which show that retrievers pre-trained on DL pairs has a distinct advantage over that on CM pairs, while combining both gives extra improvement.

Negative Passage In practice, negative sampling

is essential for learning a high-quality encoder. Besides in-batch negative, our reported HLP employs one additional negative for each query. We further explore the impact of the additional negatives during pre-training. In our ablation study, pre-training with additional negatives improves the results significantly, which may be attributed to using more in-batch pairs for text matching. More details on implementation and negative sampling strategies can be found in Appendix B.

6.2 Analysis on Q-P Overlap

We carry out extensive analysis on the Q-P lexical overlap in the task of retrieval. Specifically, we tokenize q, p using the BERT tokenizer and measure the Q-P overlap as the proportion of the question tokens that appear in the corresponding passage. Based on the degree of Q-P overlap, we divided the NQ dev set into five categories for further analysis.

Distribution of Q-P Overlap Figure 3 shows both the pre-training and the retrieved pairs of HLP have a more similar overlap distribution with the downstream NQ dataset than the other methods, which implies the consistency between the relevance provided by HLP and that in real information-seeking scenario.

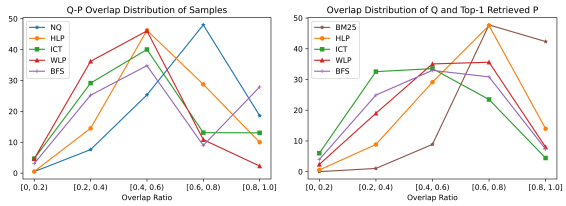

Figure 3: Distribution of overlap on pseudo and downstream Q-P pairs (left), and that between the query and the top-1 passage retrieved by different pre-trained models (right).

Retrieval Performance vs. Q-P Overlap Figure 4 shows the top-20 retrieval accuracy on the samples with varying degrees of Q-P overlap. Both figures show that the retrievers are more likely to return answer-containing passages when there is higher Q-P overlap, suggesting that all these models can exploit lexical overlap for passage retrieval. Under the zero-shot setting, HLP outperforms all the methods except BM25 when r is larger than 0.8, which reflects the strong reasoning ability of HLP and the overlap-dependent nature of the term-based retrievers. After fine-tuning, models with additional pre-training perform better than the vanilla DPR while HLP outperforms all other methods in most of the cases. It is important to note that HLP is pre-trained on more high-overlap text pairs while it performs better than all the other methods when fewer overlaps are provided. We speculate that this is because the overlapping in HLP Q-P pairs mostly comes from the factual information, such as entity, which introduces fewer superficial cues, allowing for better adaptation to the downstream cases.

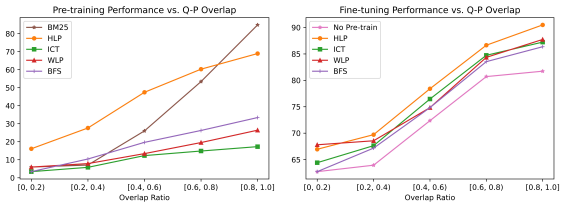

Figure 4: Top-20 retrieval accuracy of pre-training (left) and fine-tuning (right) on the divided NQ dev set.

6.3 Human Evaluation on Q-P pairs

We conduct human evaluation to investigate the proportion of Q-P pairs that convey the similar fact-level information. Specially, we randomly selected one hundred examples from our constructed Q-P pairs and asked annotators to identify whether the query and the corresponding passage convey similar facts. Each case is evaluated by three annotators and the result is determined by their votes. Our results are shown in Table 6, and we further present case studies in Appendix G.

	DL	CM	WLP
Votes	61%	40%	15%

Table 6: Human evaluation on pseudo Q-P pairs constructed by different methods.

7 Conclusion

This paper proposes Hyperlink-induced Pre-training (HLP), a pre-training method for OpenQA passage retrieval by leveraging the online textual relevance induced by hyperlink-based topology. Our experiments show that HLP gains significant improvements across multiple QA datasets under different scenarios, consistently outperforming other pre-training methods. Our method provides insights into OpenQA passage retrieval by analyzing the underlying bi-text relevance. Future work involves addressing tasks like MS MARCO where the granularity of the information-seeking target is at the passage level.

References

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.

Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A neural click model for web search. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 531–541.

Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. *arXiv preprint arXiv:2002.03932*.

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.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 65–74.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR.

Mandar Joshi, Eunsol Choi, Daniel S 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.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Patrick S. H. Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *CoRR*, abs/2102.07033.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith B. Hall, and Ryan T. McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1075–1088.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

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. In *CoCo@ NIPS*.

Barlas Oğuz, Kushal Lakhotia, Anchit Gupta, Patrick Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Wen-tau Yih, Sonal Gupta, et al. 2021. Domain-matched pre-training tasks for dense retrieval. *arXiv preprint arXiv:2107.13602*.

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.

Revanth Gangi Reddy, Vikas Yadav, Md Arafat Sultan, Martin Franz, Vittorio Castelli, Heng Ji, and Avirup Sil. 2021. Towards robust neural retrieval

models with synthetic pre-training. *arXiv preprint arXiv:2104.07800*.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Devendra Singh Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers for open-domain question answering. *arXiv preprint arXiv:2101.00408*.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2021. Can question generation debias question answering models? a case study on question-context lexical overlap. *arXiv preprint arXiv:2109.11256*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

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

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Parameter Details

For the pre-training, all models including our reproduced baselines are trained with 20 million Q-P pairs with in-batch negative sampling. Specifically, our reported HLP is trained on the combination of 10 million DL pairs and 10 million CM pairs. The HLP (DL) and HLP (CM) reported in Table 4 are trained on 10 million DL pairs and 10 million CM pairs, respectively. More parameters details are shown in the table below.

Hyperparameter	Pre-training	Fine-tuning
Epoch	5	40
Batch Size	400	256
GPU Resource	32GB GPU × 8	32GB GPU × 8
Learning Rate	2e-5	2e-5
Warmup Ratio	0.1	0.1
Learning Rate Decay	Linear	Linear
Shared Encoder	True	False
Maximum Q Length	150	256
Maximum P Length	256	256

B Negative Sampling

While negative sampling plays an import role in contrast learning, we have explored different types of negatives to pair with queries: (1) Random negatives: passages randomly selected from the corpus (2) Overlap negatives: passages have entity overlap with queries but fail to match either DL or CM topology. Our experimental results in Table 7 show that the model perform better when it adopts random negatives. We conjecture that the overlap negatives may be too hard for the self-supervised pre-training. Thus, we pair one random negative to each query during pre-training.

Negative	NQ			TriviaQA		
Type	top5	top20	top100	top5	top20	top100
None	45.7	66.0	79.9	62.6	75.2	83.0
Random	**51.2**	**70.2**	**82.0**	**65.9**	**76.9**	**84.0**
Overlap	49.7	67.8	80.3	63.1	75.1	83.0

Table 7: Top-k zero-shot retrieval accuracy on NQ and TriviaQA test set of HLP model pre-trained with different types of negatives.

C Data Analysis on NQ Samples

In this part, we detailedly discuss how we conduct exploratory data analysis on NQ training set to determine the hyperlink-based topology. Driven by a strong interest in what kind of roles the overlapping spans play between the queries q and passages p, we conduct exploratory data analysis on the widely-used NQ dataset.

Specifically, we recognize all entities and mentions from the queries and the passages using TagMe (Ferragina and Scaiella, 2010) for further investigation. As a result, we observe about 55% queries q either explicitly mentions the titles of p or the successfully links to the document where p originated via TagMe, which motivates us to construct the dual-link topology where the pseudo queries q mention p via a hypertext. Moreover, we observe about 45% queries do not mention title of q but share the same mentions with p which encourages us to adopt the co-mention topology where the pseudo q and p both mention a third-party document through hypertext.

D Fact-level Evidence Reduction

Intuitively, we assume the mentioned entity, let's say e_Y mentioned in a Wikipedia document X, is used to describe the topical entity e_X of this document. In other words, e_Y is likely to attend

in a topically relevant fact or event, which can be represented as a triple <e_X, r_{XY}, e_Y> where r_{XY} is a latent relation between e_X and e_Y.

Given any passage pair (q, p) from Wikipedia, we call q and p have fact-level evidence if they both entail a fact that can be represented as a triple <e_X, r_{XY}, e_Y> where e_X, e_Y are entities and r_{XY} is their corresponding relations. Further, if both passages q and p contain representative hypertext or topic of e_X and e_Y, we say this fact-level evidence can be induced by hyperlink-based topology, namely hyperlink-induced fact. In this session, we prove that any Q-P pair with hyperlink-induced fact while satisfying answer containing is within either DL or CM hyperlink-based topology.

Following the example above, both q and p contain a factual triple <e_X, r_{XY}, e_Y>. Since mentioned entities are used to describe the topical entity, we have facts <e_Q, r_{QX}, e_X>, <e_Q, r_{QY}, e_Y> at q-side while <e_P, r_{PX}, e_X>, <e_P, r_{PY}, e_Y> at p-side. Further, p contains <e_P, r_{PQ}, e_Q> because of the answer containing condition.

Case1: The entity $e_P = e_X$ or $e_P = e_Y$. Then q provides facts <e_Q, r_{QP}, e_P> where r_{QP} is likely but not necessarily to be identical to r_{PQ} in p. In this case, (q, p) fits in the Dual-link topology in our definition.

Case2: The entity $e_P \neq e_X$ and $e_P \neq e_Y$. Then given the facts <e_Q, r_{QX}, e_X> at q-side, and <e_P, r_{PX}, e_X>, <e_P, r_{PQ}, e_Q> at p-side, (q, p) fits in the Co-mention topology.

E Pseudo Code for HLP Pairs

Algorithm 1: HLP Pairs Identification

Notation:

$q, p \leftarrow$ Wikipedia passages
$t_Q \leftarrow$ Topical entity of passage q
$\mathcal{M}(q) \leftarrow$ The set of entities mentioned in q
$d_{in}(q) \leftarrow$ in-degree of the Wikipedia entity t_Q
$K \leftarrow$ in-degree threshold for CM pairs

```
Def IsDL(q, p):
    if t_P ∈ M(q) & t_Q ∈ M(p) then
        return 1
    else
        return 0
;
Def IsCM(q, p):
    foreach m ∈ M(q) do
        if d_in(m) < K & m ∈ M(p) &
        t_Q ∈ M(p) then
            return 1
        else
            return 0
```

F Case Studies on Multi-hop Retrieval

We evaluate HLP on multi-hop scenario where knowledge from different documents need to be associated for retrieval. Besides significant improvements shown in Table 5, we conduct case study to investigate its capability on knowledge-intensive retrieval. In Table 8, a complex question is proposed, requiring the retriever firstly to retrieve the document *"Apple Remote"* and then *"Front Row (software)"* to fetch the final answer. Our HLP successfully retrieves both golds in the top-10 retrieved passages while the vanilla DPR fails. We find 6 items retrieved by HIS are related to the brand *"Apple"* while 4 are by DPR, showing stronger comprehension and associative ability from HLP.

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?
Evidence Passage: 1. Apple Remote: The Apple Remote is a remote control device ... was originally designed to interact with the Front Row media ... 2. Front Row (software): Front Row is a discontinued media ... is controlled by an Apple Remote or the keyboard function keys ...
Top-10 Retrieved Titles (Ours): Apple Remote; ITunes Remote; Remote computer; Wii Remote; Remote Shell; Kinect; Siri Remote; Front Row (software); Apple TV; Spinning pinwheel;
Top-10 Retrieved Titles (DPR's): Apple Remote; ITunes Remote; Console (video game CLI); Control Panel (Windows); Apple Wireless Keyboard; Chooser (Mac OS); Remote computer; Media player (software); ToggleKeys; Button (computing);

Table 8: Case studies on HotpotQA dataset. Blue colored text denote the titles of gold passages while red is the final answer span.

G Case Studies on Q-P Paraphrase

Besides human evaluation, we present case studies on HLP Q-P pairs, which is shown in Table 9 and Table 10. As we can see in the tables, a few lexical variants of entities and fact-level paraphrasing are presented across questions and passages, which can be interpreted as factual evidence for OpenQA passage matching. For example, entity-level variants such as *"Robert and Richard Sherman"* vs. *"Sherman Brothers"*, and fact-level paraphrases such as *"Abby Kelley and Stephen Symonds Foster ... working for abolitionism"* vs. *"... radical abolitionists, Abby Kelley Foster and her husband Stephen S. Foster"* can be found in our examples.

Query	Passage
Title: **Abby Kelley** Liberty Farm in Worcester, Massachusetts, the home of Abby Kelley and Stephen Symonds Foster, was designated a National Historic Landmark because of its association with their lives of working for abolitionism.	Title: **Worcester, Massachusetts** Two of the nation's most radical abolitionists, Abby Kelley Foster and her husband Stephen S. Foster, adopted Worcester as their home, as did Thomas Wentworth Higginson, the editor of The Atlantic Monthly and Emily Dickinson's avuncular correspondent, and Unitarian minister Rev. Edward Everett Hale. The area was already home to Lucy Stone, Eli Thayer, and Samuel May Jr. They were joined in their political activities by networks of related Quaker families such as the Earles and the Chases, whose organizing efforts were crucial to ...
Title: **Callisto Corporation** They were best known for their series of computer games for the Macintosh in the 1990s, including ClockWerx, Spin Doctor, Super Maze Wars and Super Mines.	Title: **ClockWerx** ClockWerx is a computer game created by Callisto Corporation that was released in 1995. The game was originally released by Callisto under the name SSpin Doctor". Later, with some game play enhancements, it was published by Spectrum HoloByte as "Clockwerx which was endorsed by Alexey Pajitnov according to the manual. A 3DO Interactive Multiplayer version was planned but never released. The object of the game is to solve a series of increasingly difficult levels by swinging a rotating wand from dot to dot until the player reaches the "goal" dot. Enemy wands ...
Title: **Sivaji Ganesan** Some of his famous hits during this period are " Vasantha Maligai ", " Gauravam ", " Thanga Pathakkam " and " Sathyam ".	Title: **Vasantha Maligai** Vasantha Maligai is a 1972 Indian Tamil -language romance film, directed by K. S. Prakash Rao and produced by D. Ramanaidu . The film stars Sivaji Ganesan and Vanisri , and is the Tamil remake of the 1971 Telugu film " Prema Nagar ". " Vasantha Maligai " was released on 29 September 1972 and became a major commercial success, running in theatres for nearly 750 days. A digitally restored version of the film was released on 8 March 2013, and another one on ...
Title: **Say Anything (band)** Around this time, the band also released " Alive with the Glory of Love " as a single.	Title: **Alive with the Glory of Love** "Alive with the Glory of Love" is the first single from Say Anything \'s second album " ...Is a Real Boy ". "Alive with the Glory of Love" was released to radio on June 20, 2006. The song was a hit for the band, charting at number twenty-eight on the Alternative Songs chart. The song, described as an "intense and oddly uplifting rocker about a relationship torn by the Holocaust," by the " Pittsburgh Post-Gazette ", is actually semi-biographical in nature, telling the story of songwriter and vocalist Max Bemis \'s ...
Title: **Dorothy Sue Hill** Hill taught home economics from 1960 to 1969 for the Allen Parish School Board and from 1969 to 1992 for the Beauregard Parish School Board .	Title: **Allen Parish School Board** Allen Parish School Board is a school district headquartered in Oberlin in Allen Parish in southwestern Louisiana , United States. From 1960 to 1969, Dorothy Sue Hill, the state representative for Allen, Beauregard , and Calcasieu parishes, taught home economics for Allen Parish schools.

Table 9: Examples of DL Q-P pairs where text in blue gives evidence and answer.

Query	Passage
Title: **Daniel Gormally** In 2015 he tied for the second place with David Howell and Nicholas Pert in the 102nd British Championship andeventually finished fourth on tiebreak.	Title: **Nicholas Pert** In 2015, Pert tied for 2nd–4th with David Howell and Daniel Gormally, finishing third on tiebreak, in the British Chess Championship and later that year, he finished runner-up in the inaugural British Knockout Championship, which was held alongside the London Chess Classic. In this latter event, Pert, who replaced Nigel Short after his late withdrawal, eliminated Jonathan Hawkins in the quarterfinals and Luke McShane in the semifinals, then he lost to David Howell 4–6 in the final.
Title: **Ojuelegba, Lagos** Ojuelegba is a suburb in Surulere local government area of Lagos State.	Title: **Simi (singer)** ... on September 8, 2017. Her third studio album " Omo Charlie Champagne, Vol. 1 " was released to coincide with her thirty-first birthday on April 19, 2019. She launched her record label Studio Brat in June 2019. Simi was born on 19 April 1988 in Ojuelegba, a suburb of Surulere, Lagos State, as the last of four children. In an interview with Juliet Ebirim of " Vanguard " newspaper, Simi revealed that her parents separated when she was 9 years old. She also revealed that she grew up as a ...
Title: **The Aristocats** Longtime Disney collaborators Robert and Richard Sherman composed multiple songs for the film, though only two made it in the finished product.	Title: **In Search of the Castaways** Later sang the Sherman Brothers \' theme song \' The Aristocats \' from Disney\'s 1970 animated film "The Aristocats". Ïn Search of the Castaways" was a commercial success. Upon its initial release, it earned $4.9 million in North American theatrical rentals. It was one of the 12 most popular movies at the British box office in 1963. " The New York Times " declared: Ït is, as we say, a whopping fable, more gimmicky than imaginative, but it doesn\'t lack for lively melodrama that is more innocent and wholesome than much of the ...
Title: **Jang Jin-young** As of 2008, Jang was one of the highest paid stars in the Korean film industry, earning in the region of per film.	Title: **Scent of Love** Scent of Love (Scent of Chrysanthemums) is a 2003 South Korean film, and the directorial debut of Lee Jeong-wook. The film is based on a novel of the same name by Kim Ha-in, and stars Jang Jin-young and Park Hae-il in the lead roles. Like her character, Jang Jin-young battled stomach cancer and died in 2009. The film received an around of 900,000 admissions nationwide and on May 16, 2003 the film was screened at the Cannes Film Festival. University student Seo In-ha meets a ...
Title: **Vera Menchik** Vera Menchik ("Vera Frantsevna Menchik" 16 February 1906 – 26 June 1944) was a Russian-born British-Czechoslovak chess player who became the first women\'s world chess champion.	Title: **Paula Wolf-Kalmar** Paula Wolf-Kalmar (11 April 1880 - 29 September 1931) was an Austrian chess master, born in Zagreb. She took 5th at Meran 1924 (unofficial European women's championship won by Helene Cotton and Edith Holloway). After the tournament three of the participants (Holloway, Cotton and Agnes Stevenson) defeated three others (Kalmar, Gulich and Pohlner) in a double-round London vs. Vienna match. She was thrice a Women's World Championship Challenger. She took 3rd, behind Vera Menchik and Katarina Beskow at London 1927 ...

Table 10: Examples of CM Q-P pairs where text in blue gives evidence while red gives answer.