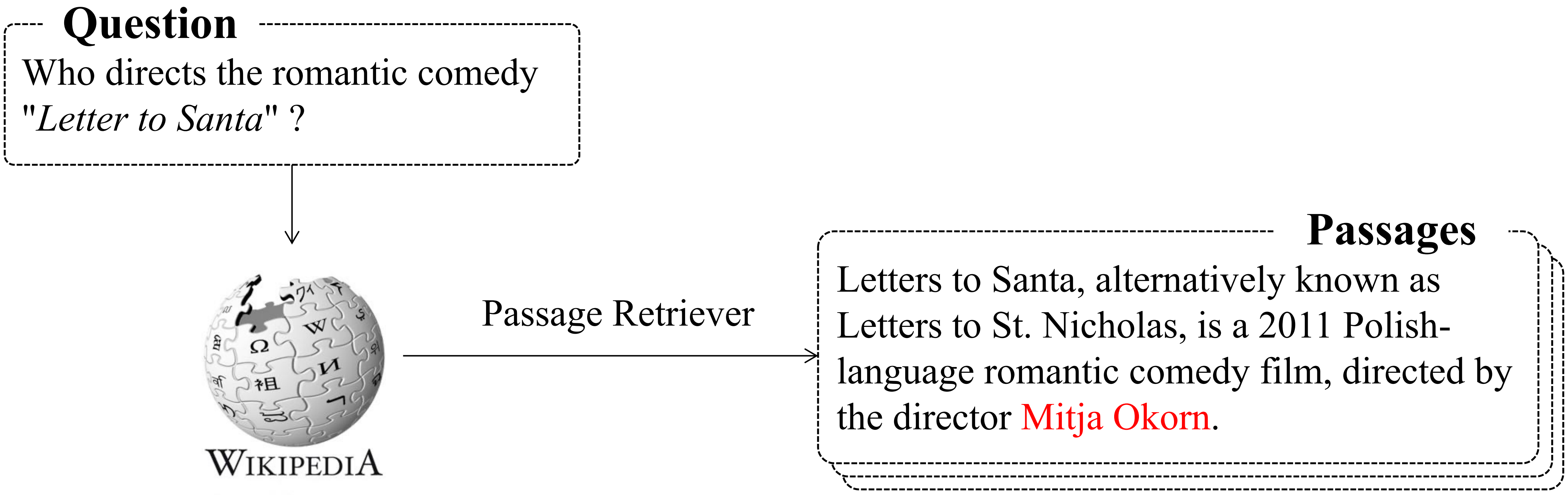
Hyperlink-induced Pre-training for Passage Retrieval in Open-domain Question Answering

Jiawei Zhou[1], Xiaoguang Li[2], Lifeng Shang[2], Lan Luo[3], Ke Zhan[3], Enrui Hu[3],
Xinyu Zhang[3], Hao Jiang[3], Zhao Cao[3], Fan Yu[3], Xin Jiang[2], Qun Liu[2], Lei Chen[1]

[1]The Hong Kong University of Science and Technology
[2]Huawei Noah's Ark Lab
[3]Distributed and Parallel Software Lab, Huawei

Background

Passage Retrieval in Open-domain QA



Question

Who directs the romantic comedy "*Letter to Santa*" ?

Passages

Letters to Santa, alternatively known as Letters to St. Nicholas, is a 2011 Polish-language romantic comedy film, directed by the director Mitja Okorn.

relevant

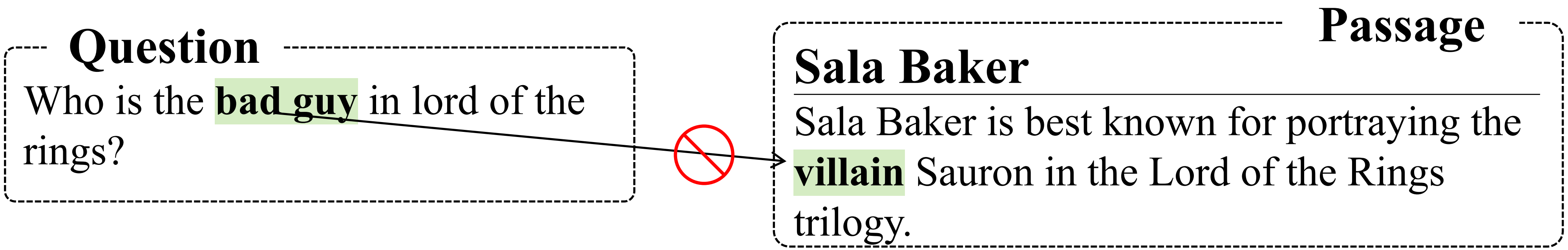
answer-containing

Previous Approaches

Sparse Representation for Retrieval

Traditional retrievers, such as TF-IDF and BM25, matches keywords efficiently based on sparse representations.

incapable when deep semantic understanding is required.



Question

Who is the **bad guy** in lord of the rings?

Passage

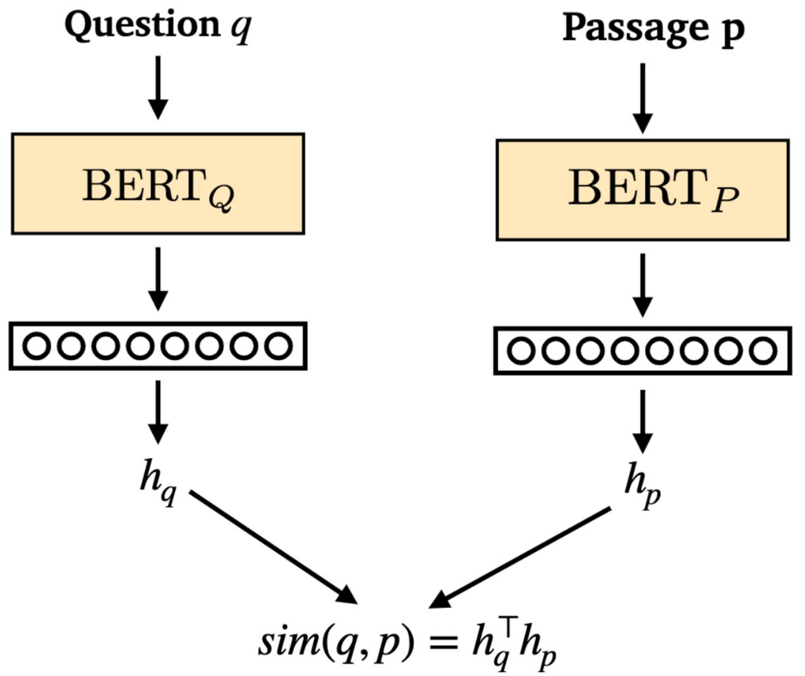
Sala Baker

Sala Baker is best known for portraying the **villain** Sauron in the Lord of the Rings trilogy.

Previous Approaches

Dense Passage Retrieval (DPR)

Dense retrievers learn **dense representations** to semantically match an embedded query to the most relevant passages.



outperform BM25

heavily rely on labelled data

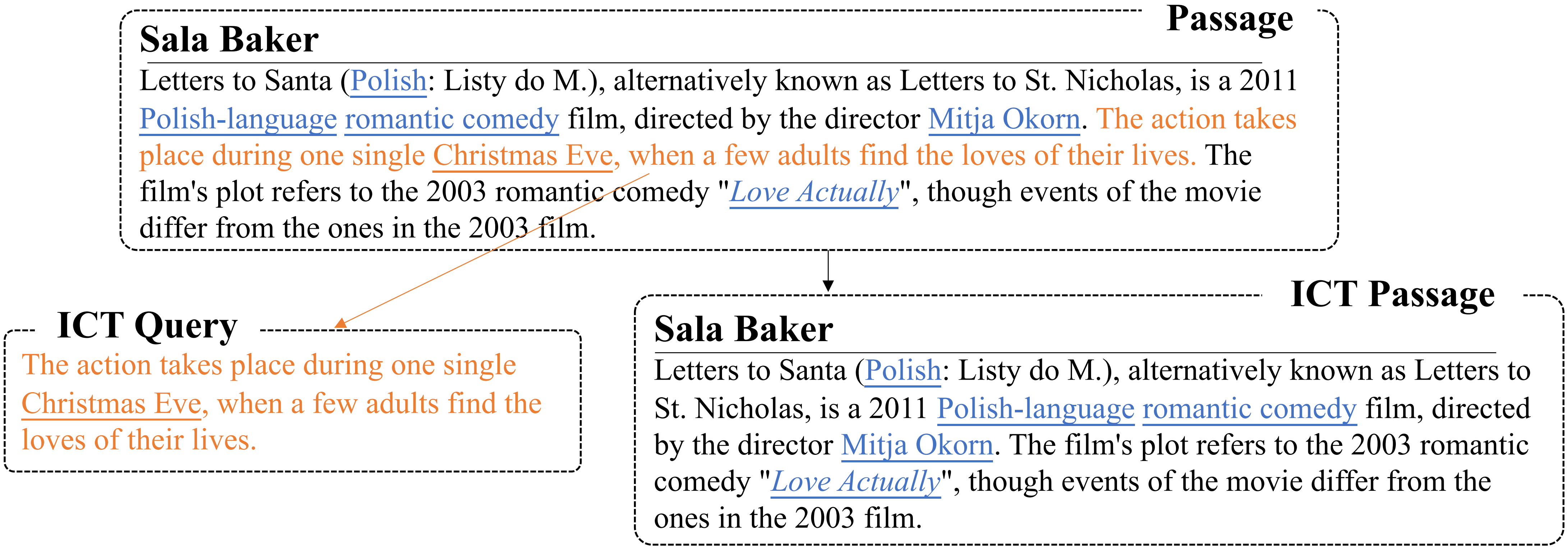
Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Previous Approaches

Dense Retriever + Pre-training

1. Pre-training on weakly supervised data
   [Lee et al., 2019; Chang et al., 2020; Guu et al., 2020; Sachan et al., 2021]

Inverse Cloze Task (ICT)



Passage

Sala Baker

Letters to Santa (Polish: Listy do M.), alternatively known as Letters to St. Nicholas, is a 2011 Polish-language romantic comedy film, directed by the director Mitja Okorn. The action takes place during one single Christmas Eve, when a few adults find the loves of their lives. The film's plot refers to the 2003 romantic comedy "*Love Actually*", though events of the movie differ from the ones in the 2003 film.

ICT Query

The action takes place during one single Christmas Eve, when a few adults find the loves of their lives.

ICT Passage

Sala Baker

Letters to Santa (Polish: Listy do M.), alternatively known as Letters to St. Nicholas, is a 2011 Polish-language romantic comedy film, directed by the director Mitja Okorn. The film's plot refers to the 2003 romantic comedy "*Love Actually*", though events of the movie differ from the ones in the 2003 film.

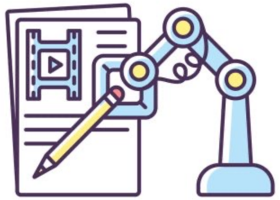
Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering

Previous Approaches

Dense Retriever + Pre-training

1. Pre-training on weakly supervised data
   [Lee et al., 2019; Chang et al., 2020; Guu et al., 2020; Sachan et al., 2021]

2. Data Augmentation via Question Generation
   [Ma et al., 2021, Reddy et al., 2021 and O˘guz et al., 2021]



tend to generate questions with high lexical overlap,

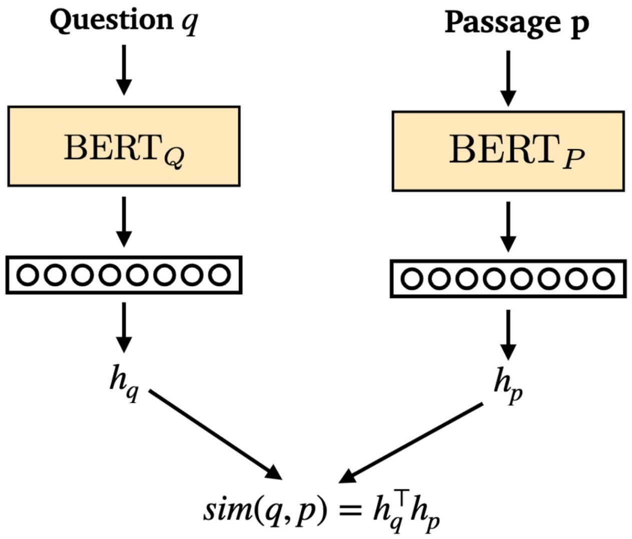
which amplifies the bias of QA dataset

Our Approach

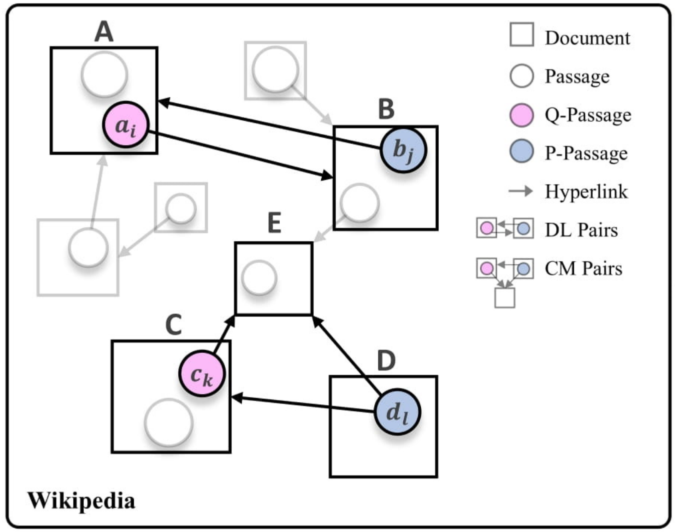
HLP: HyperLink-induced Pre-training

Pre-training a dense passage retriever on the supervision signal derived from Wikipedia hyperlinks

DPR



Wikipedia



Motivation

Why leverage Hyperlink 🔗?

- The purpose of hyperlinks is similar to that of retrievers -- namely, to help users seek relevant information.

Main Motivation

- Wikipedia naturally contains a large number of hyperlinks.

Quantity guaranteed

- These hyperlinks have been widely used and updated by the community.

Quality guaranteed

Prior knowledge

Q-P relevance in OpenQA

*What kind of **relevance** should exist between query and passage?*

1. Evidence Existence

   Evidence, such as entities and their corresponding relations, should exist across the query and the targeted passage.

2. Answer Containing

   The targeted passage should contain the information-seeking target (i.e., the answer) of the query.

Q-P relevance in HLP

*What kind of **relevance** provided in HLP pseudo Q-P pairs?*

1. Evidence Existence in HLP

   Co-occurrence of entities that presented as hypertext or topics in q and p.

query passage

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

entity-level factual information conveyed by the context

2. Answer Containing in HLP

   We consider the topical entity of document Q as the information-seeking target of q. In this case, the targeted passage p should mention t_Q.

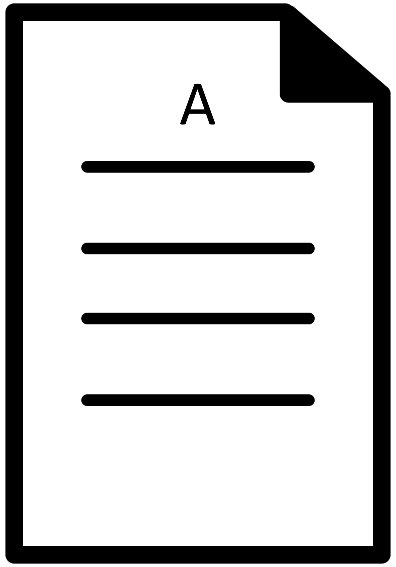
$$t_Q \subseteq p$$

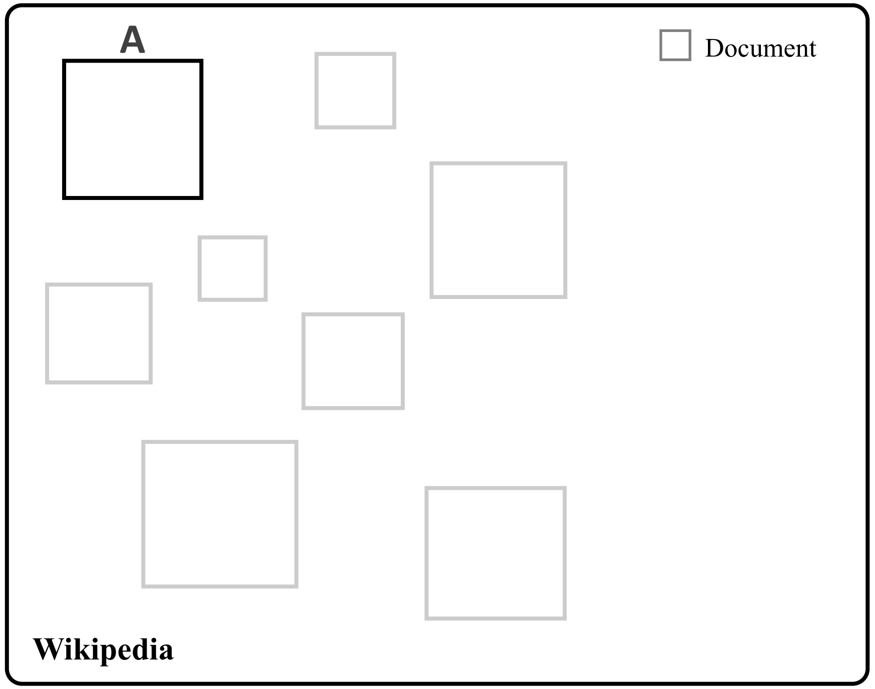
topical entity of document Q and any query q originated from Q

Our Approach

HLP: HyperLink-induced Pre-training

Given a Wikipedia document



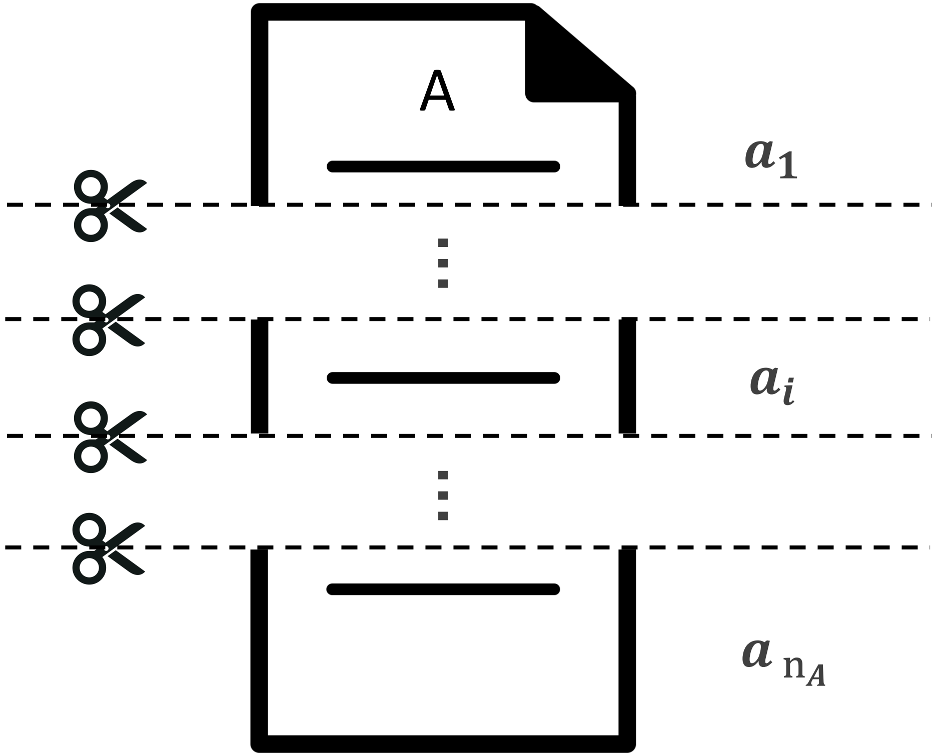


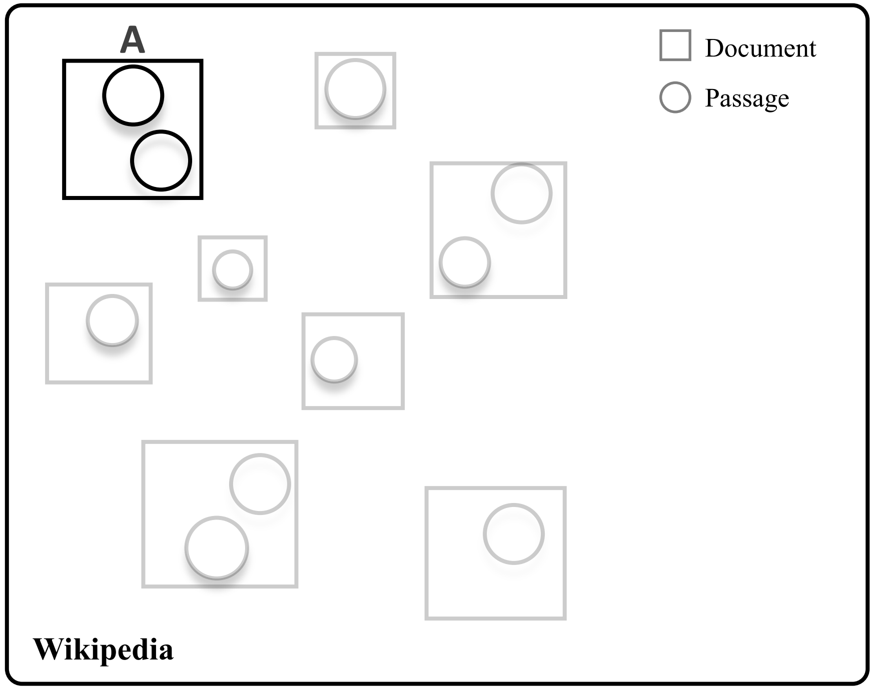
Our Approach

HLP: HyperLink-induced Pre-training

Split the documents into disjoint chunks as passages

$$A = [a_1, \dots, a_i, \dots, a_{n_A}]$$

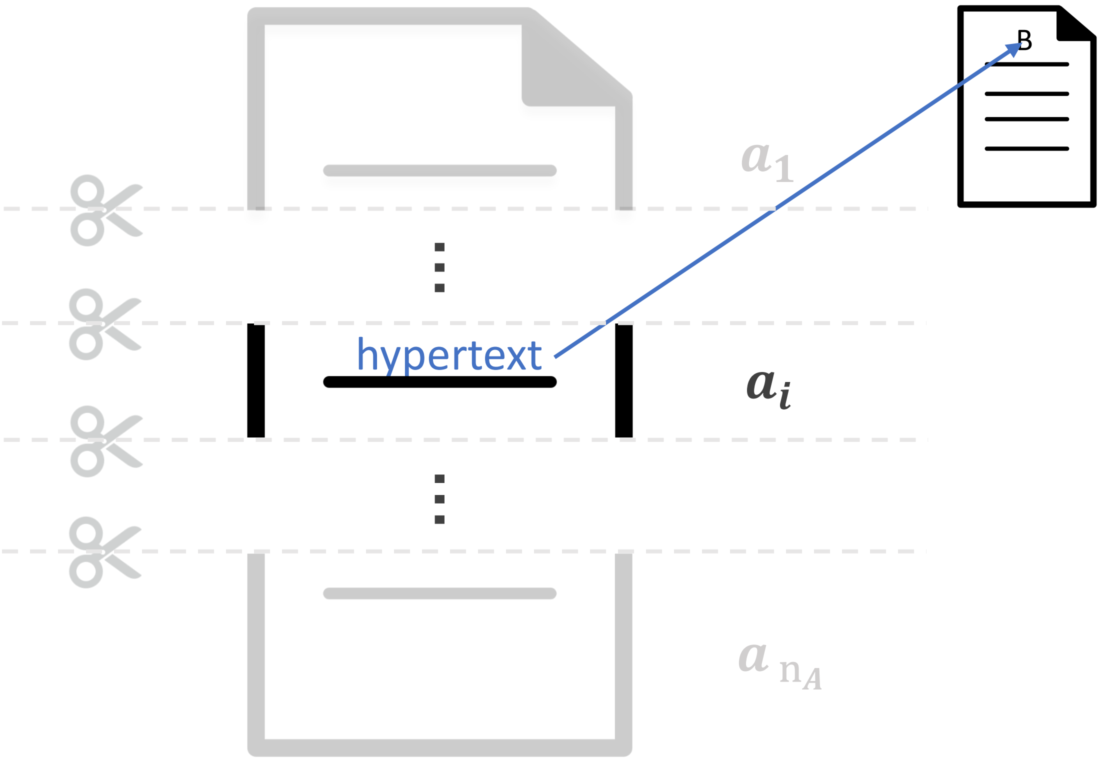


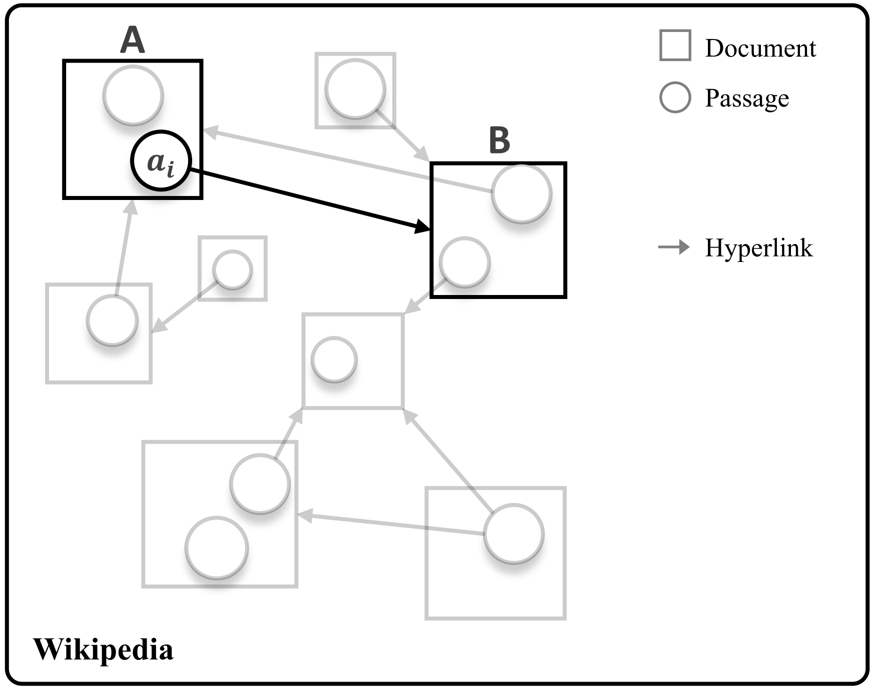


Our Approach

HLP: HyperLink-induced Pre-training

Take passages as nodes and hyperlinks as links to construct a Wikipedia graph





Our Approach

HLP: HyperLink-induced Pre-training

We propose two kinds of hyperlink topologies
where we extract the pseudo Q-P pairs for pretraining

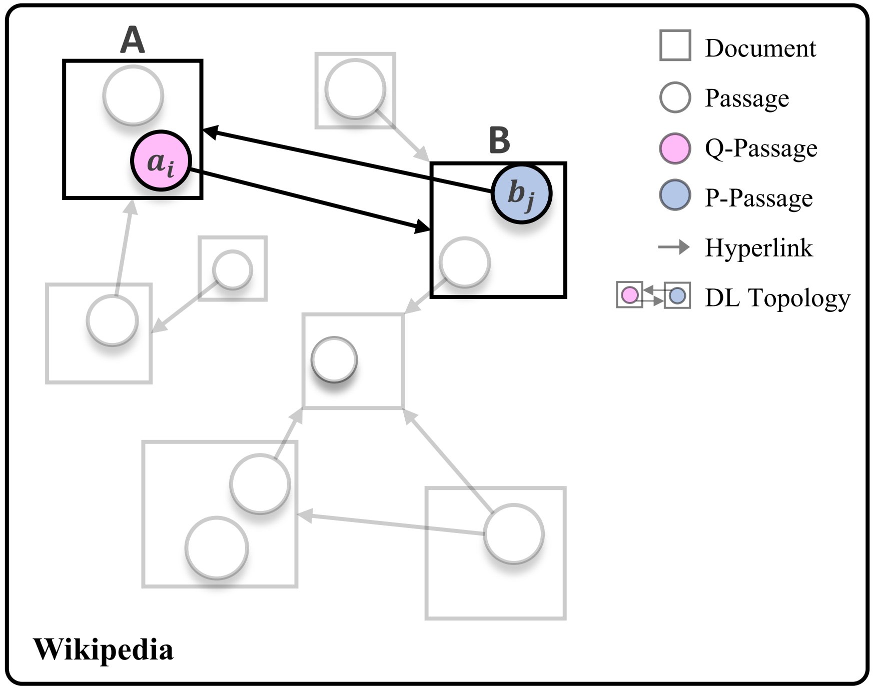
1. Dual-link example as (a_i, b_j)
   A passage pair (q, p) if they link to each other.

a_i, b_j mentions the topical entity of the other

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

b_j contains m_A which is a_i 's
information-seeking target as we assume

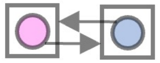
$$t_A \approx m_A \quad \text{and} \quad m_A \subseteq b_j$$ Answer ✓

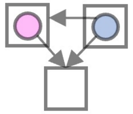


Our Approach

HLP: HyperLink-induced Pre-training

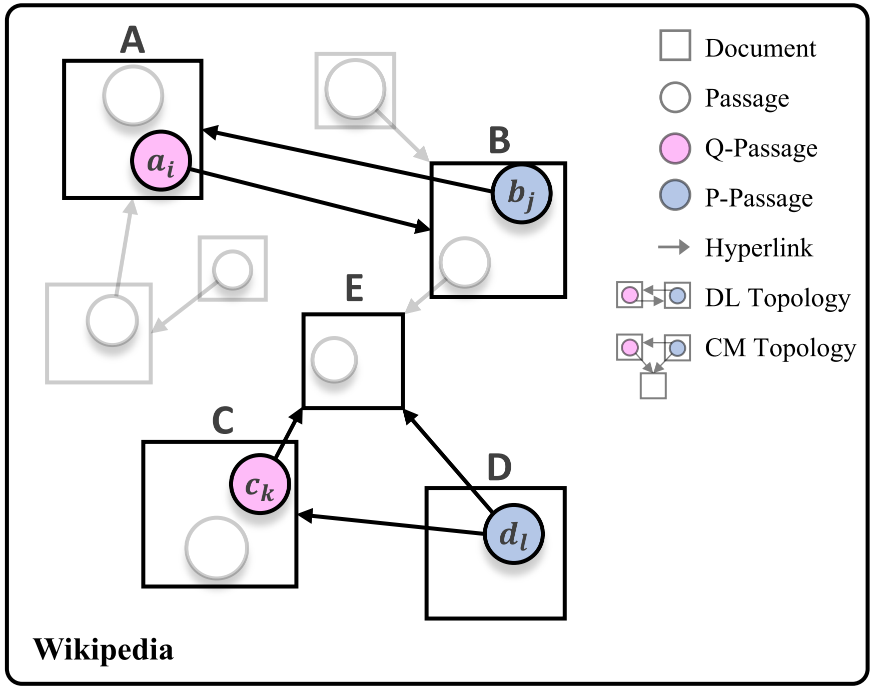
We propose two kinds of hyperlink topologies where we extract the pseudo Q-P pairs for pretraining

1. Dual-link  example as (a_i, b_j)
   A passage pair (q, p) if they link to each other.

2. Co-mention  example as (c_k, d_l)
   A passage pair (q, p) if they both link to a third-party document and p links to q

$\{e_C, e_E\} \subseteq \mathcal{F}_{(c_k)} \cap \mathcal{F}_{(d_l)}$ **Evidence** ✓

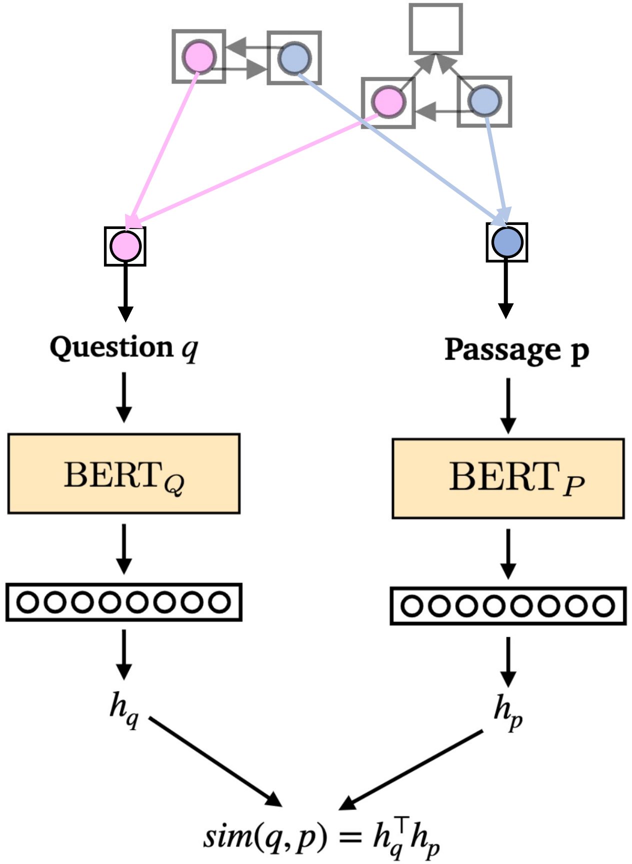
$t_C \approx m_C$ and $m_C \subseteq d_l$ **Answer** ✓

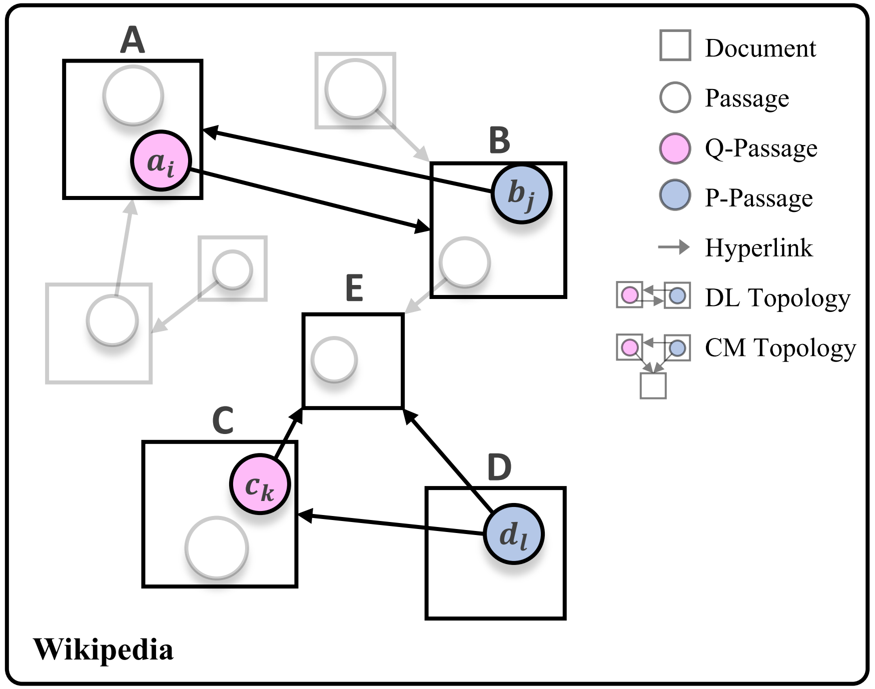


Our Approach

HLP: HyperLink-induced Pre-training

HLP is a dense passage retriever pre-trained on 20m DL and CM pairs.





Our Approach

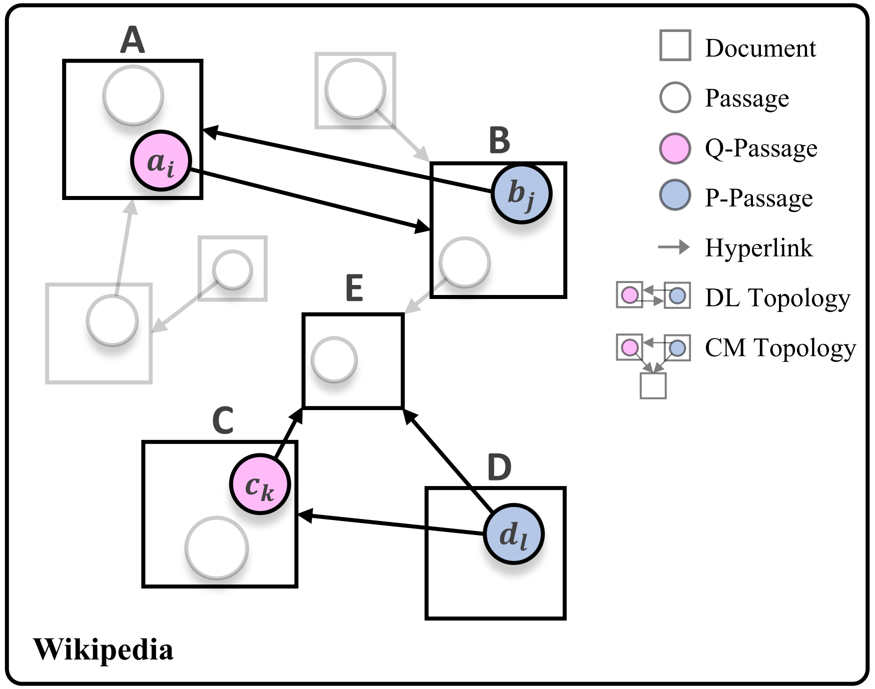
HLP: HyperLink-induced Pre-training

Why does HLP work?

1. Similar Q-P relevance between that in downstream and that induced by hyperlinks.

2. The construction of HLP Q-P pairs does NOT rely on lexical overlap but the guidance of hyperlink.

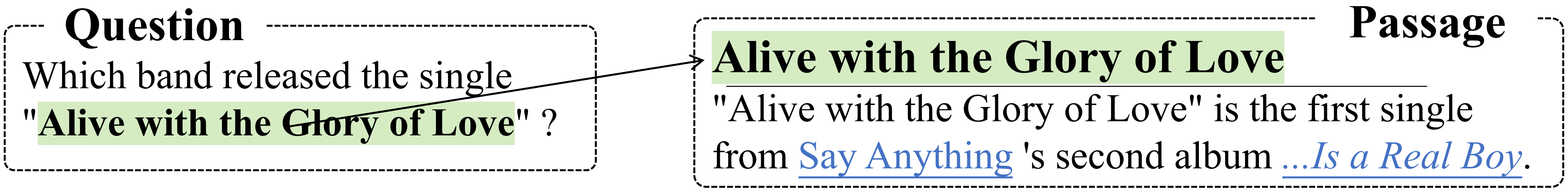
introducing more semantic similarity and lexical diversity.



Why DL & CM topologies?

We have conducted analysis on the downstream dataset (NQ), and found:

- 55% *q* mention the topical entity of *p* or successfully link to the golden document by the entity linking tool.



Question

Which band released the single "**Alive with the Glory of Love**" ?

Passage

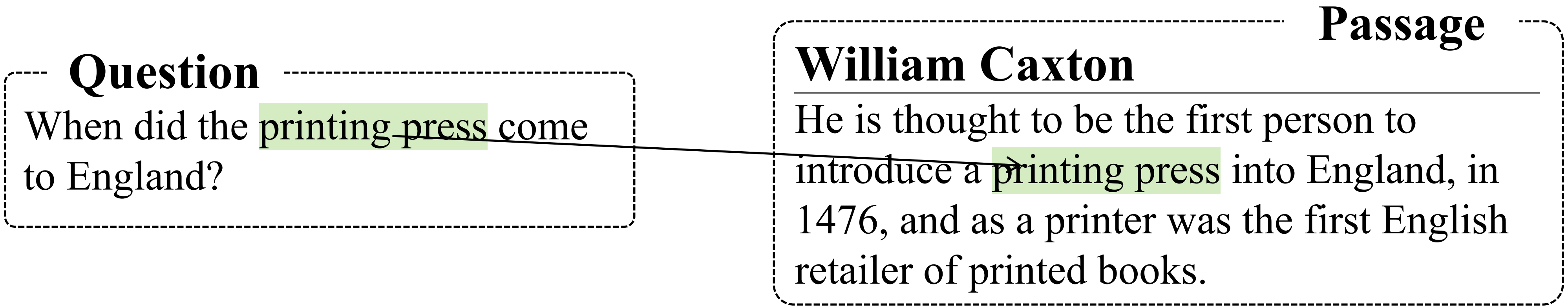
Alive with the Glory of Love

"Alive with the Glory of Love" is the first single from Say Anything 's second album *...Is a Real Boy*.

Why DL & CM topologies?

We have conducted analysis on the downstream dataset (NQ), and found:

- 55% *q* mention the topical entity of *p* or successfully link to the golden document by the entity linking tool.
- 45% *q* share same mentions with *p*.



Question

When did the printing press come to England?

Passage

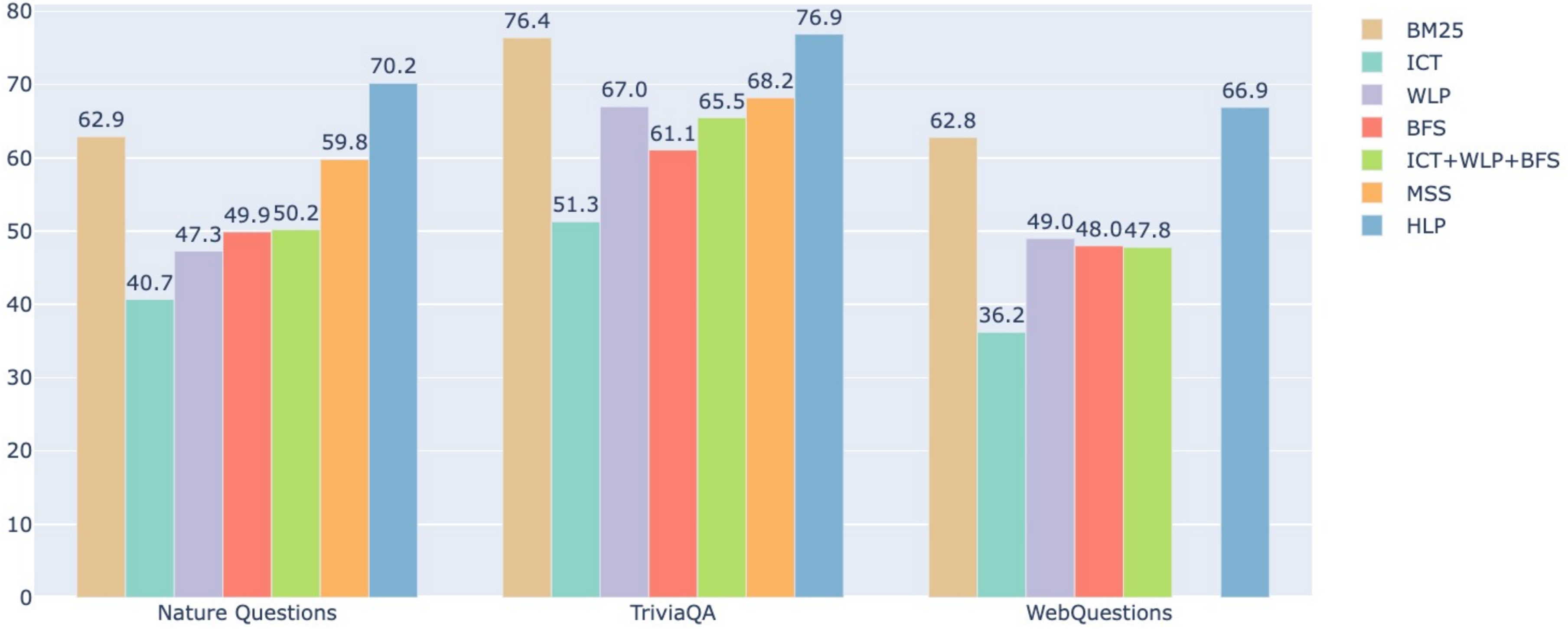
William Caxton

He is thought to be the first person to introduce a printing press into England, in 1476, and as a printer was the first English retailer of printed books.

Experiment #1

Passage Retrieval (Main Result)

Top-20 (zero-shot) retrieval accuracy after pre-training:



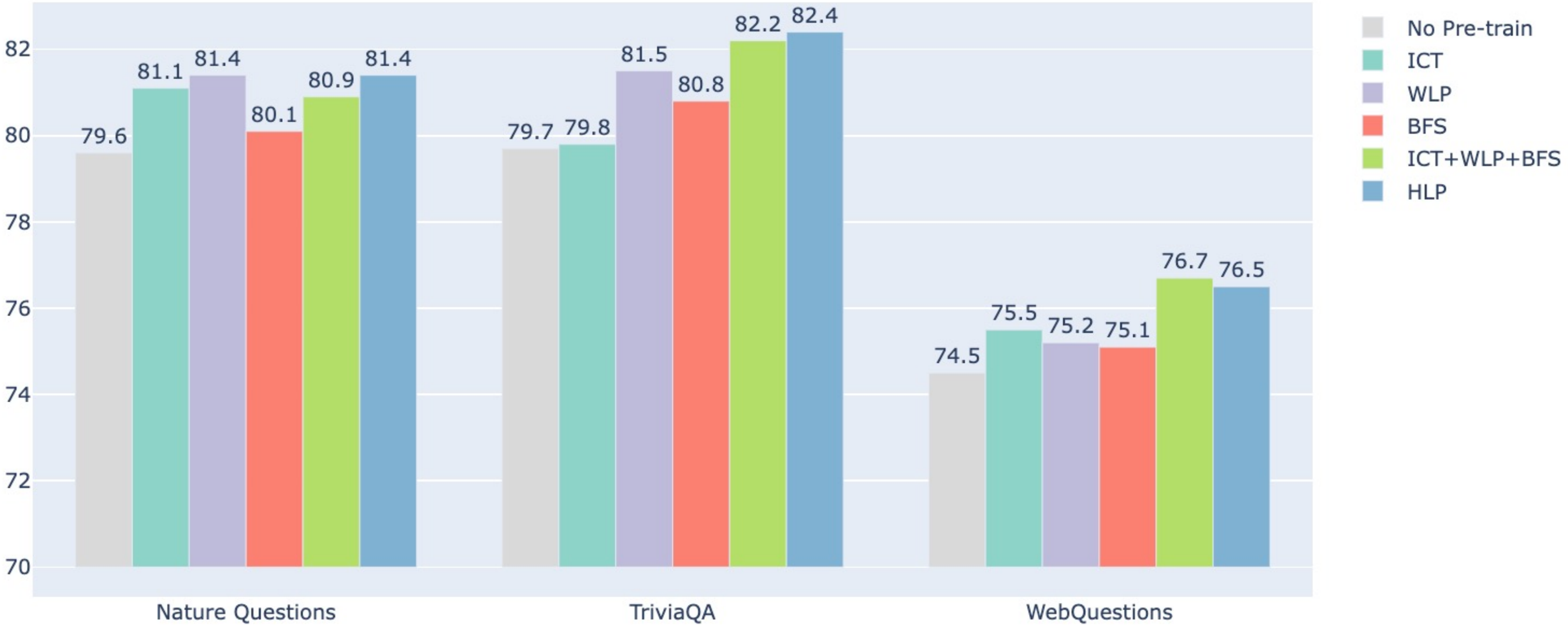
~4% higher than BM25

~20% higher than other pre-training methods

Experiment #1

Passage Retrieval (Main Result)

Top-20 retrieval accuracy after fine-tuning:



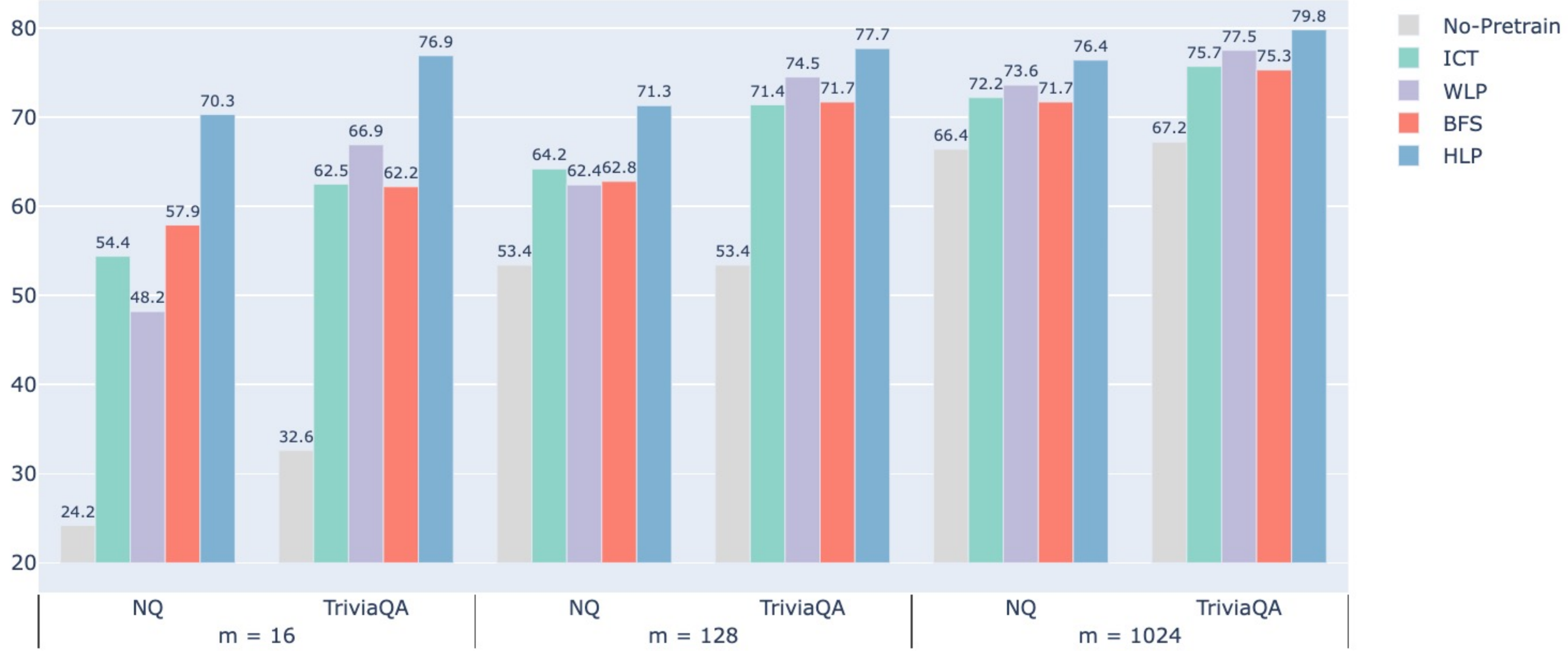
~2% higher than that without pre-train

~1% higher than other pre-trained retrievers

Experiment #2

Few-shot Learning

Top-20 retrieval accuracy after fine-tuning on *m* samples:



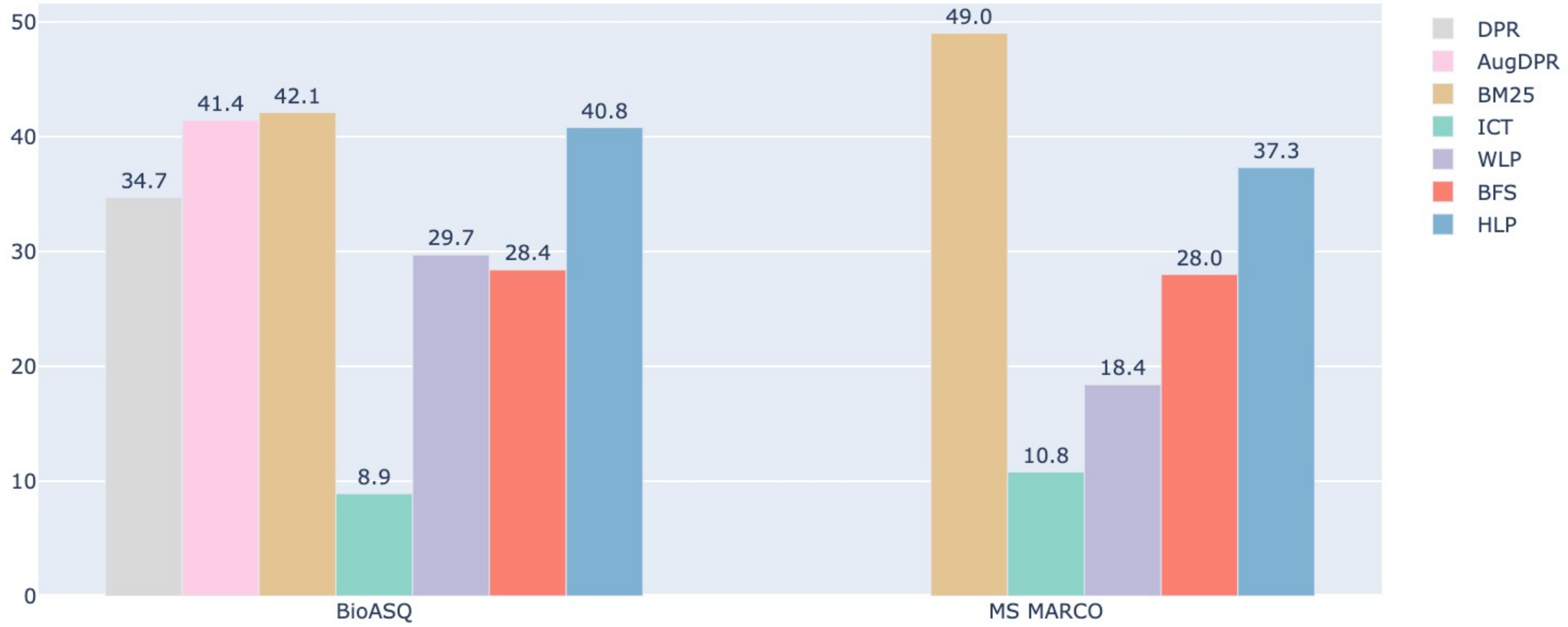
- Intermediate pre-training give significant improvement under few-shot scenario.
- HLP outperforms the others by a larger margin especially when *m* is smaller

Experiment #3

Out-of-domain (OOD) Scenario

Non-Wikipedia Corpus

Top-20 (zero-shot) retrieval accuracy:



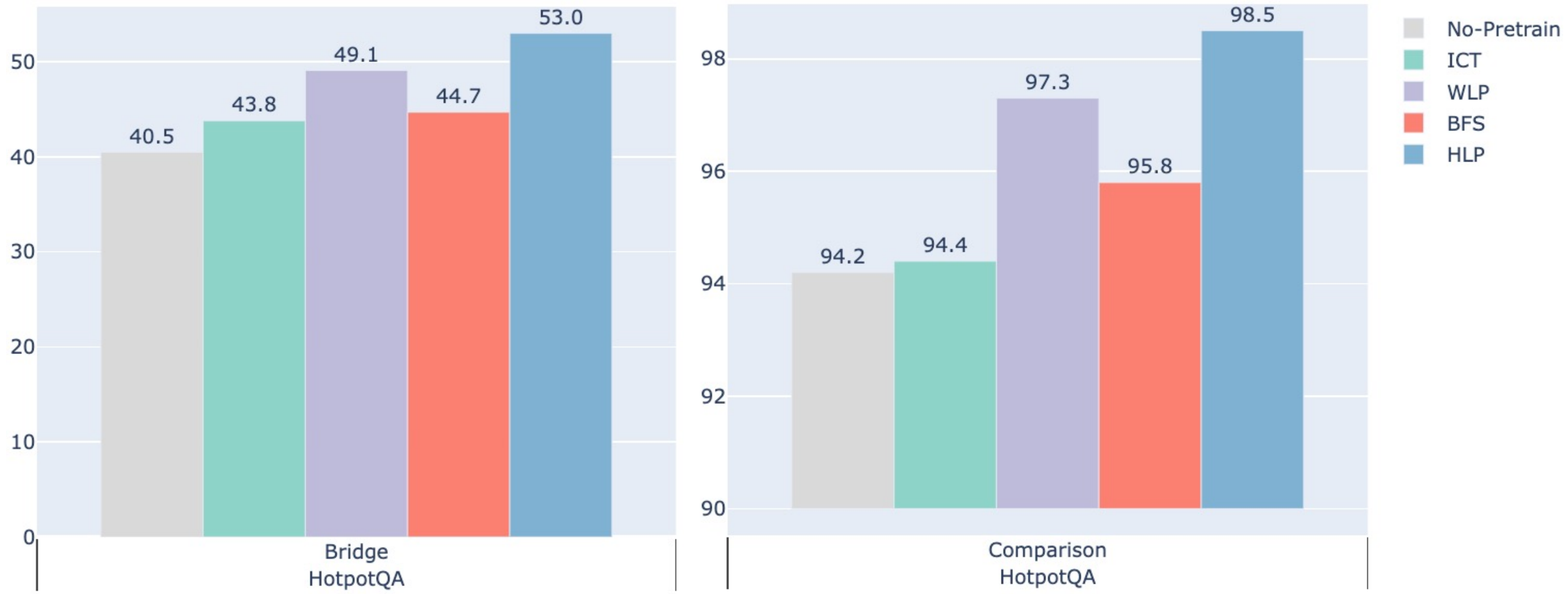
- HLP significantly outperforms ICT, WLP and BFS on both datasets.
- HLP matches BM25 and AugDPR on BioASQ.
- HLP falls behind BM25 on MS MARCO for two reasons:
  1. higher Q-P overlap observed in MS MARCO
  2. information seeking target of MS MARCO is passage rather than a text span

AugDPR has access to NQ labeled data
while HLP is trained in unsupervised data

Experiment #4

Multi-hop Retrieval

Top-20 retrieval accuracy:



~7% higher than other pre-training baselines on bridge questions
~2% higher than other pre-training baselines on comparison questions

Analysis #1

Ablation Study

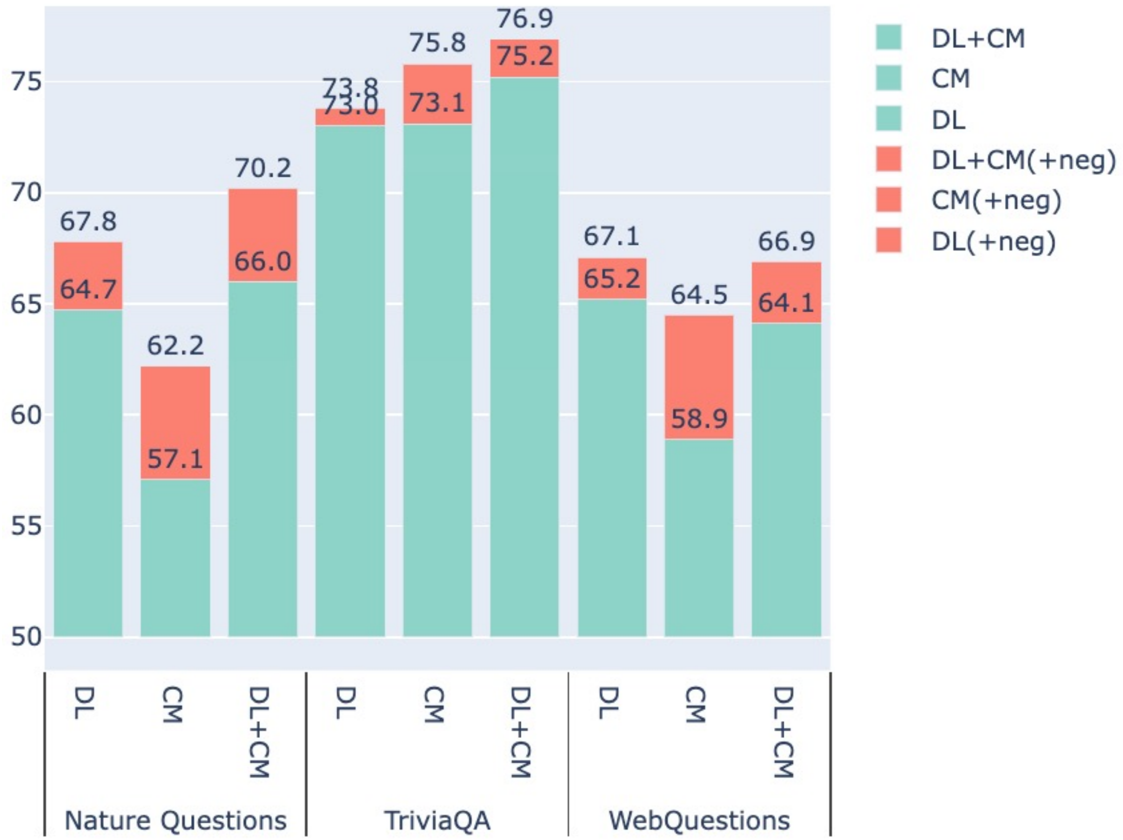
Topologies:

- DL mostly outperforms CM
- Combining DL and CM makes better performance

Negatives:

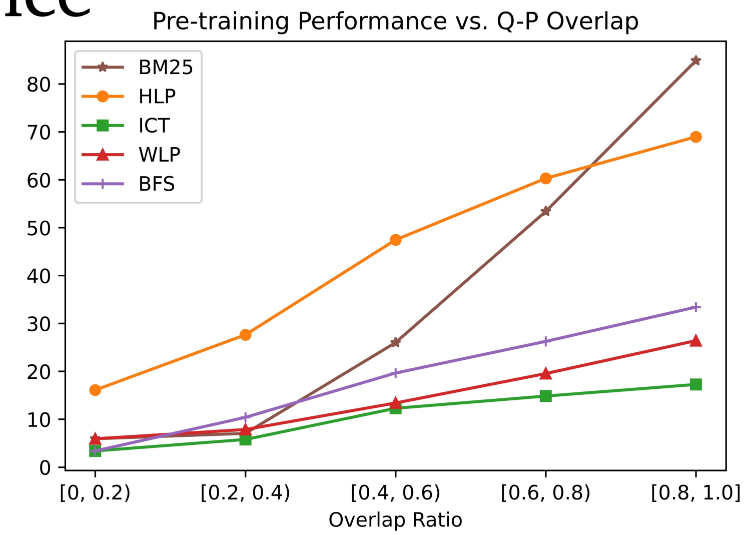
- Employing one additional negatives per query can significantly improve the result

more passages for contrastive learning



Analysis #2

Q-P Overlap v.s. Performance


Pre-training Performance vs. Q-P Overlap

Overlap v.s. Retrieval Performance:

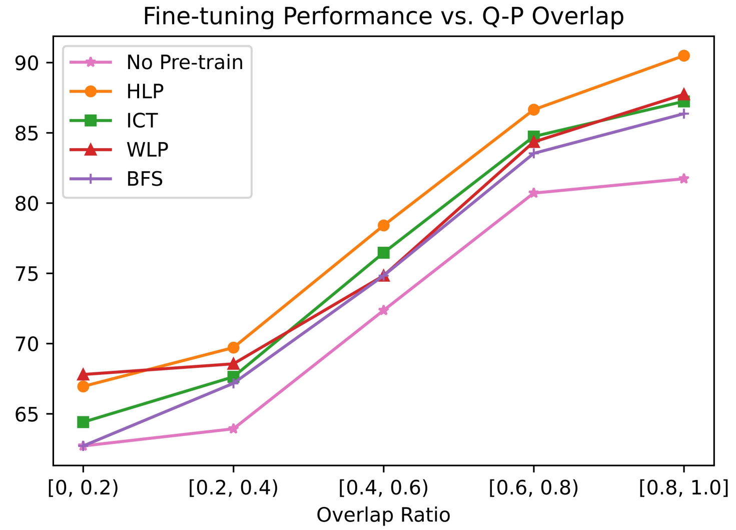
- With the higher overlap, the models are more likely to retrieve the ground truth passage.

superficial lexical overlap is easy to capture

- HLP outperforms all pre-training methods in zero-shot settings regardless of how much overlap there is.

capable of deep semantic understanding

- Except when the Q-P overlap is extremely high (>0.8), HLP can significantly outperform BM25 without any fine-tuning.

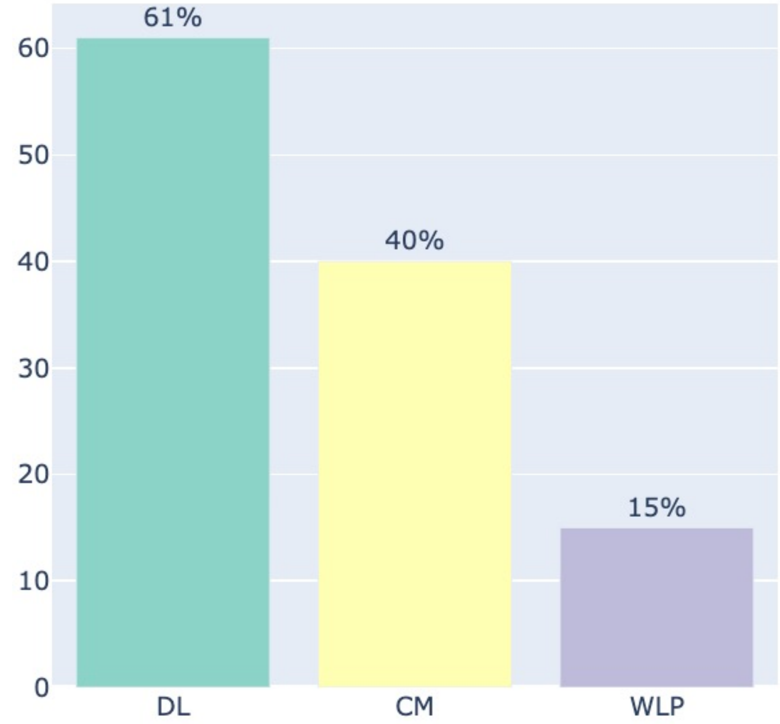

Fine-tuning Performance vs. Q-P Overlap

Analysis #3

Human Evaluation on Paraphrasing

Annotators are asked to identify whether the query and the passage are paraphrases (i.e., conveying similar facts).

HLP introduces more semantic similarity



Q & A

Code: https://github.com/jzhoubu/HLP

Email: jzhoubu@ust.hk