**HLP：由网页超链接驱动的开放领域问答检索预训练**

**总述：**

本文提出了一种新的预训练方法，称为由超链接驱动的预训练（Hyperlink-induced Pre-training, HLP），这种方法的核心思想：网页中的超链接往往指向与当前页面内容相关的其他页面，这种链接关系隐含了页面之间的语义联系，通过分析这些链接，可以构建出更符合人类信息检索习惯的查询-段落对，从而为预训练检索器提供更丰富的上下文信息。该方法利用网页文档中自然存在的超链接结构来增强文本之间的相关性。通过分析维基百科等大型百科中的超链接，利用超链接结构中的双链接（Dual-link）和共同提及（Co-mention）两种拓扑结构构建出更符合真实问答场景的伪问题-段落对（Q-P pairs），这些伪对在预训练过程中帮助模型学习到更深层次的文本匹配能力。

**输入与输出：**

**输入：**预训练语料库，通常是大型Web文档集合比如维基百科

**中间产物：**原文档段落间在具有双链接（Dual-link）和共同提及（Co-mention）结构的伪问题-段落对

**输出：**预训练后的检索器模型，经过HLP方法训练的检索器模型能够理解查询和文档段落之间的语义关联，并在给定查询时，从大规模文档集合中检索出相关的段落。

**详细步骤：**

**步骤1：**数据准备，选取预训练语料库并进行段落划分。本文使用的是2021年3月1日的英文维基百科的快照，先用WikiExtractor去除页面中的HTML标签、导航链接、图片、表格等非文本元素，只保留纯文本内容。之后过滤掉空白文本或标题少于三个字母的文档,将剩余的文档分割成100词左右的段落，以这些段落为节点，以超链接为边构建图。如图1所示

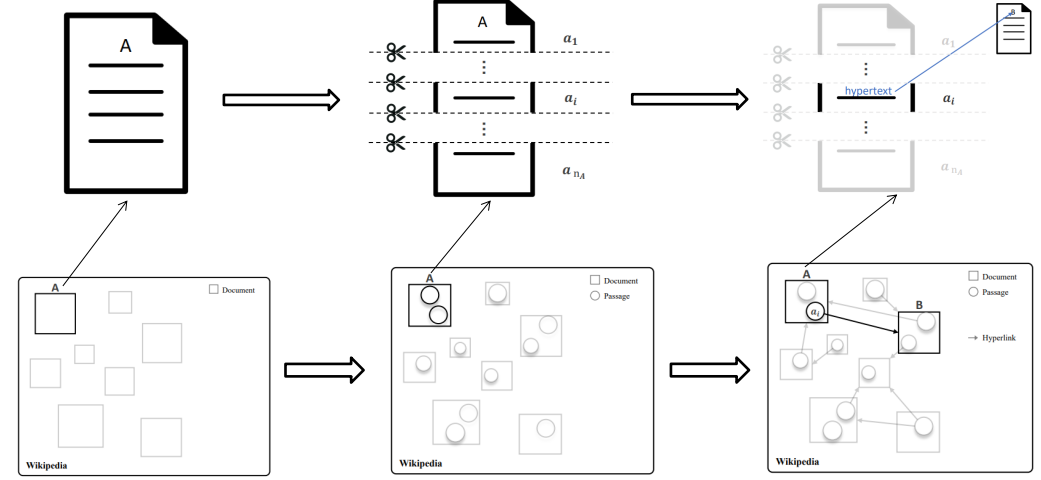


图 1 对文档划分段落并构建图

**步骤2：**超链接结构分析，分析预训练语料库中的超链接，识别出双链接（Dual-link）和共同提及（Co-mention）这两种拓扑结构，构造伪Q-P对(对于双链接结构，找到相互引用的页面对；对于共同提及结构，找到同时链接到第三方页面的页面对),如下图2所示（）为双链接对，（）为共同提及对。

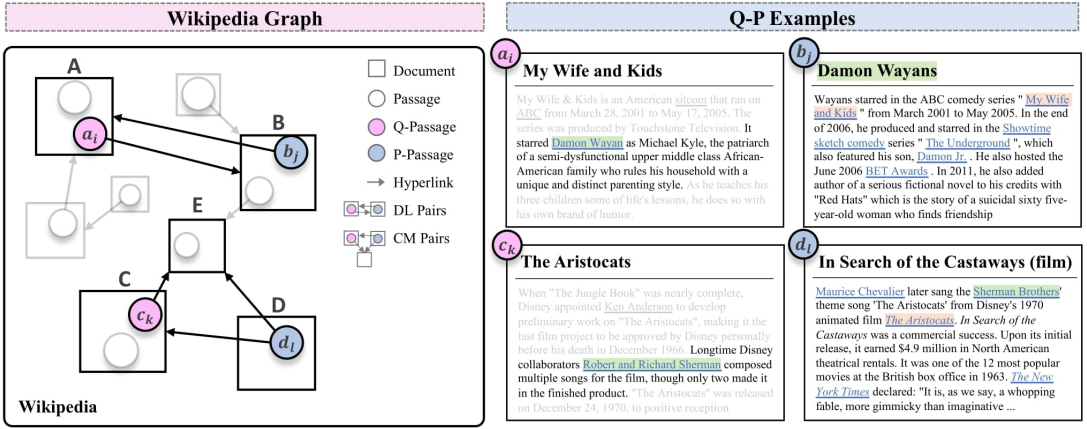


图 2 双链接对和共同提及对

**步骤3：**预训练模型构建，使用基于BERT的双编码器模型对查询和段落进行编码，采用负对数似然损失函数进行训练。

**步骤4：**模型评估与微调，在预训练完成后，评估模型在不同数据集上的性能，如零样本、少样本和跨领域场景。在特定的问答数据集上对模型进行微调，以适应实际的问答任务需求。

**案例：**

HLP(Hyperlink-induced Pre-training),由双链接或共同提及拓扑结构构造Q-P对。

ICT(Inverse Cloze Task)，随机取段落中的某句子作为query，余下段落作为passage。

WLP(Wiki Link Prediction),选取给定passage的外链接文档中的第一节的句子作为query。

由人类进行查询时提出了一个问题：“Who directed the romantic comedy 'Letters to Santa'?”由HLP、ICT、WLP方法构建的伪Q-P对如图3所示，HLP方法构造的伪Q-P对更接近人类提出的问题与正确答案。

HLP通过利用维基百科中的超链接结构来生成伪Q-P对，这些对更接近用户在互联网上实际检索时遇到的情况，并且HLP方法特别强调了事实证据的存在，即查询和目标段落之间共享的实体和关系，这种事实层面的关联性有助于提高检索器在理解查询意图和相关段落内容之间的语义关联方面的能力，另外HLP方法通过深入挖掘超链接结构，能够捕捉到更丰富的语义信息。这使得模型在处理需要深度语义理解的复杂查询时，能够提供更准确的检索结果。相比之下，ICT和WLP虽然也提供了一定程度的上下文信息或者页面级别的信息，但ICT和WLP可能更多地依赖于句子层面的上下文，不如HLP方法那样直接模拟真实世界的链接关系。

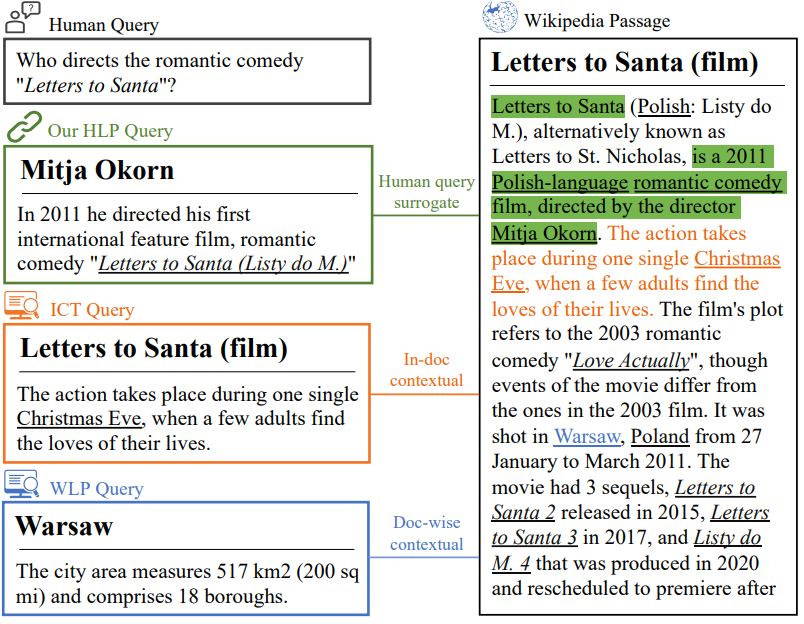


图 3 HLP、ICT、WLP方法构建的伪Q-P对

表 1 双链接Q-P 对的示例，其中蓝色文本提供证据和答案。

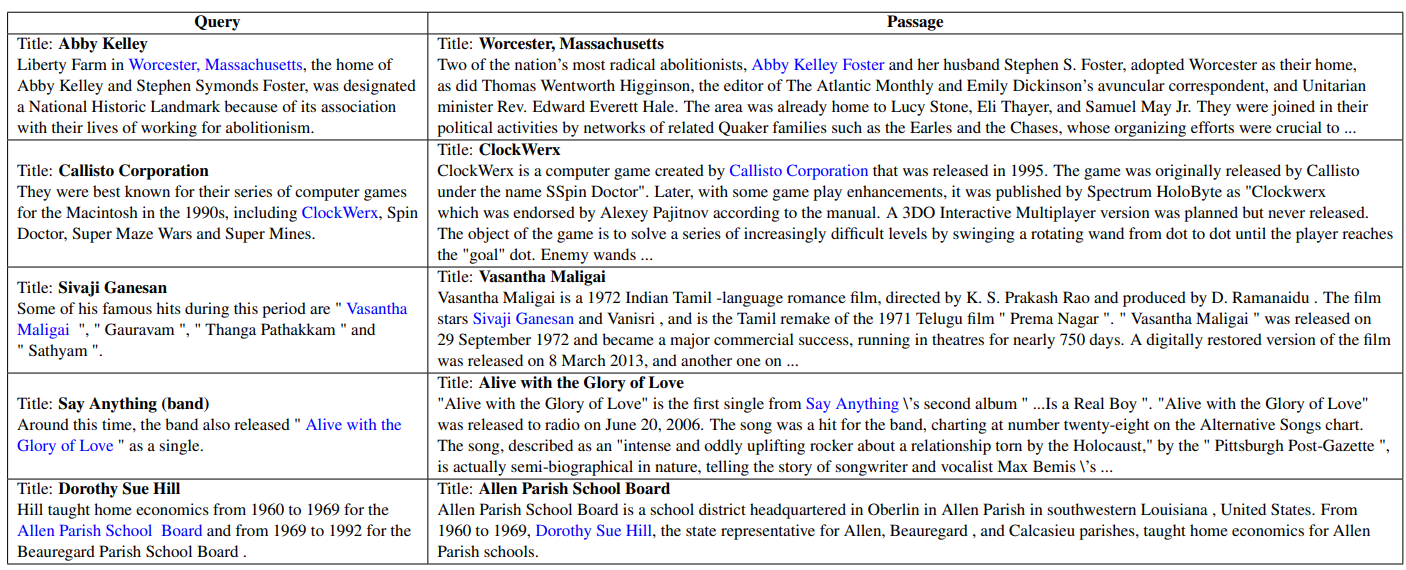
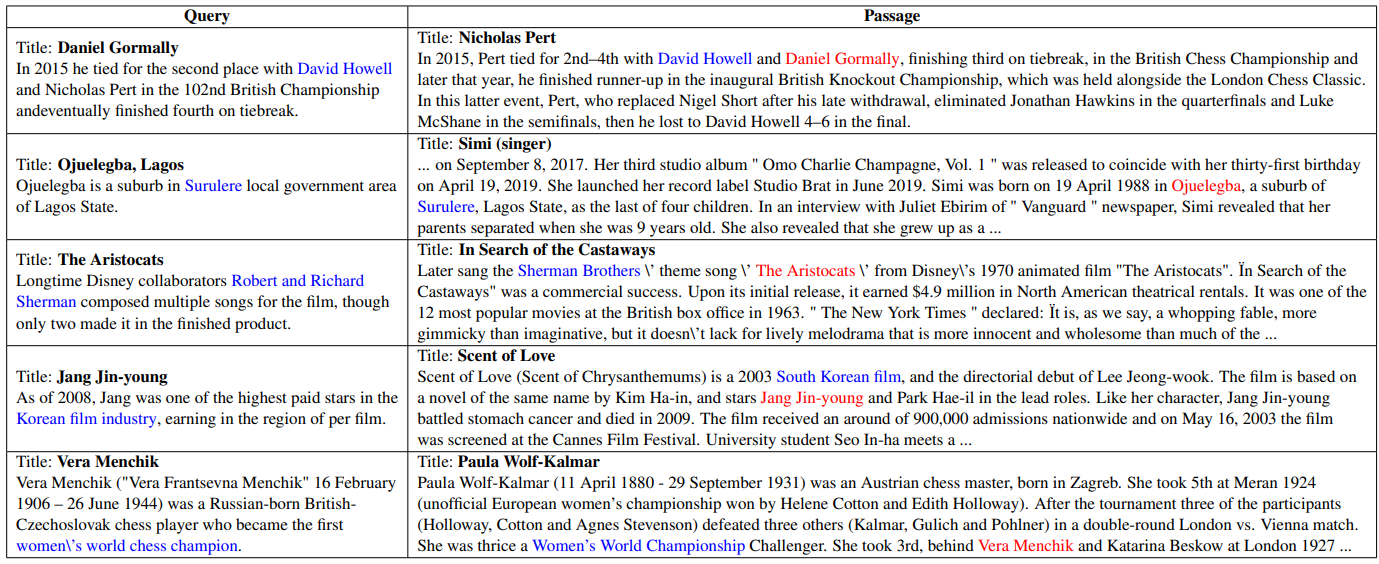


表 2 共同提及Q-P 对示例，其中蓝色文本提供证据，红色文本提供答案。



**参考文献**

1. Zhou J , Li X , Shang L ,et al.Hyperlink-induced Pre-training for Passage Retrieval in Open-domain Question Answering[J]. 2022.DOI:10.48550/arXiv.2203.06942.
2. 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, Florence, Italy. Association for Computational Linguistics.
3. Suchin Gururangan, Ana Marasovic, Swabha ´ Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics
4. 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 opendomain 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
5. 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 opendomain 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
6. Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer opendomain questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.
7. Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for opendomain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.
8. 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.