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

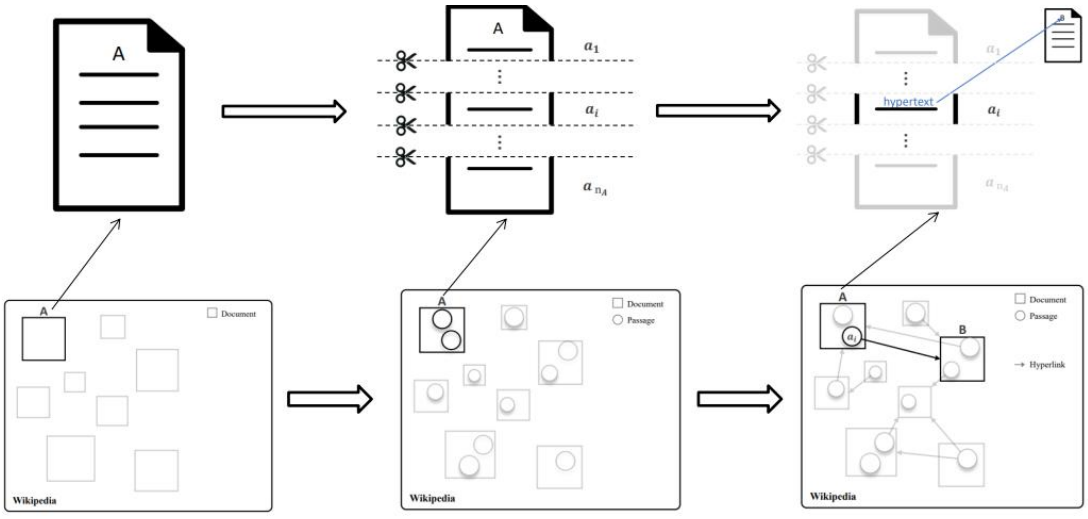

图 1 对文档划分段落并构建图

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（a_i, b_j）为双链接对，（c_k, d_l）为共同提及对。

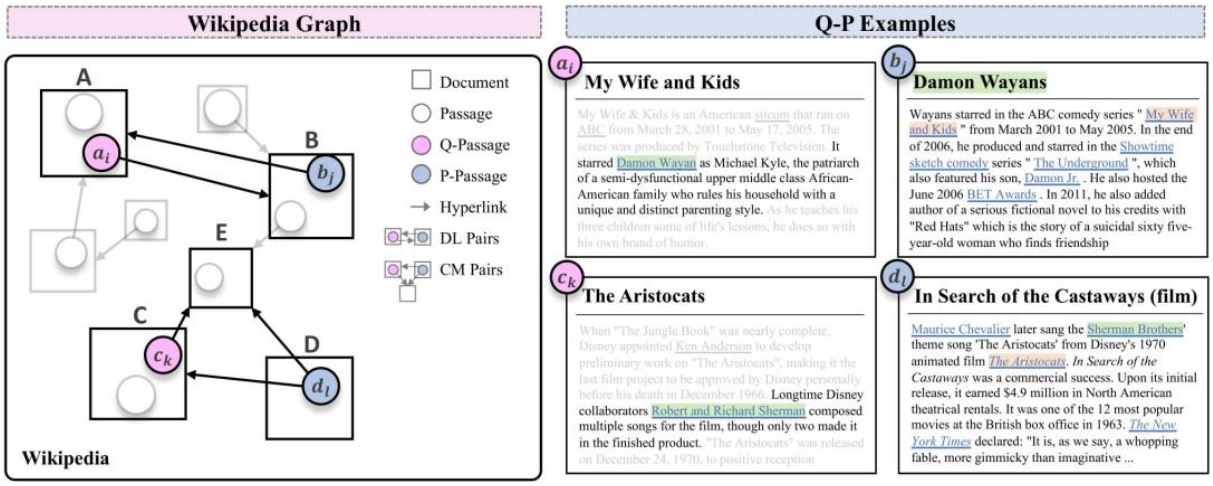

图 2 双链接对和共同提及对

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。

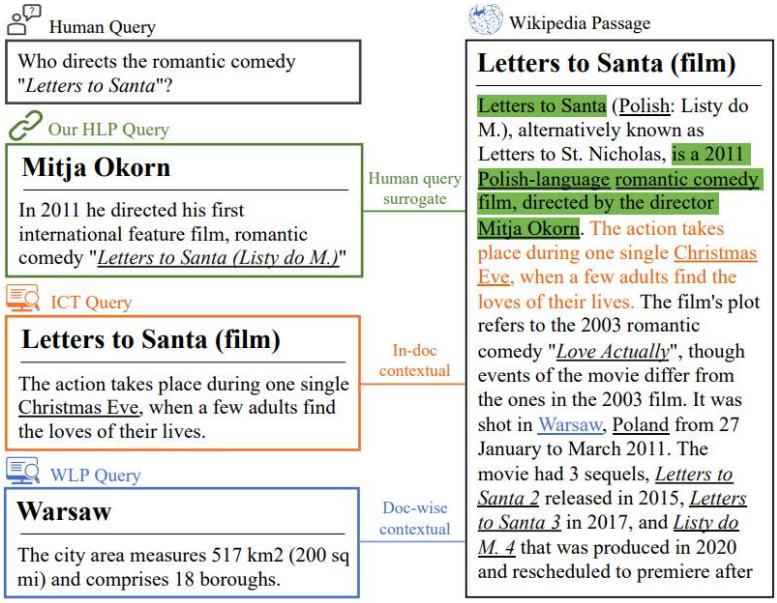

图 3 HLP、ICT、WLP 方法构建的伪 Q-P 对

表 1 双链接 Q-P 对的示例，其中蓝色文本提供证据和答案。

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.

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。

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". In 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: It 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 ...

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.