PHƯƠNG PHÁP TRUY XUẤT ZERO-SHOT DỰA TRÊN QUAN HỆ CHO BÀI TOÁN KHÁM PHÁ TRÍCH DẪN KHOA HỌC

Đào Xuân Phúc[1*], Nguyễn Thống Nhất[1], Nguyễn Thanh Xuân[2]
[*]Tác giả liên hệ, mail: phucdx@hou.edu.vn. ORCID: 0009-0000-6217-8603

Ngày tòa soạn nhận được bài báo: 15/01/2026
Ngày phản biện đánh giá: 17/03/2026
Ngày bài báo được duyệt đăng: 14/04/2026

DOI: 10.59266/houjs.2026.1169

Tóm tắt: *Nhiệm vụ chia sẻ về khám phá trích dẫn tập trung vào việc dự đoán trích dẫn chính xác từ một tập hợp ứng viên cho một đoạn văn bản đầu vào. Những thách thức chính bắt nguồn từ độ dài của các đoạn tóm tắt và độ tương đồng cao giữa các ứng viên, gây khó khăn cho việc xác định chính xác bài báo cần trích dẫn. Để giải quyết vấn đề này, bài báo trình bày một hệ thống thực hiện truy xuất top-k bản tóm tắt tương đồng nhất dựa trên các đặc trưng quan hệ được trích xuất từ đoạn văn bản đã cho. Tận dụng mô hình ngôn ngữ lớn để xác định chính xác trích dẫn phù hợp nhất. Dùng để đánh giá khung hệ thống dựa trên tập dữ liệu chuẩn của bài toán, qua đó chứng minh tính hiệu quả của phương pháp trong việc dự đoán trích dẫn.*

Từ khóa: *khám phá trích dẫn, trích xuất quan hệ, truy xuất văn bản, mô hình ngôn ngữ lớn, trí tuệ nhân tạo tạo sinh*

I. Đặt vấn đề

Dự đoán trích dẫn là một nhiệm vụ quan trọng với nhiều ứng dụng thực tiễn cho các nhà nghiên cứu. Khi bắt đầu làm việc với một vấn đề nghiên cứu, các nhà khoa học thường cần khảo sát các tài liệu khoa học liên quan đến vấn đề đó. Điều này giúp họ hiểu được hiện trạng nghiên cứu, xác định những hạn chế trong các công trình hiện có và phát triển các ý tưởng mới để cải thiện những vấn đề đó. Tuy nhiên, với sự gia tăng nhanh chóng về số lượng ấn phẩm khoa học, việc tìm kiếm chính xác các tài liệu thực sự liên quan đến mối quan tâm của họ ngày càng trở nên khó khăn. Một vấn đề nghiên cứu có thể có nhiều khía cạnh, và mặc dù có thể có nhiều cách tiếp cận để giải quyết, nhưng không phải tất cả chúng đều hữu ích cho các nhà nghiên cứu trong công việc hiện tại của họ (Hassan, 2023).

[1] Khoa Điện - Điện Tử, Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam
[2] Trường Đại học Kỹ thuật - Hậu cần Công an Nhân dân, Hà Nội, Việt Nam

Thêm vào đó, khi viết một tài liệu khoa học, các nhà nghiên cứu cần trích dẫn các nguồn cung cấp thông tin, ý tưởng và quan điểm cho họ. Quá trình này đòi hỏi sự chính xác để đảm bảo rằng người đọc có thể truy vết và xác minh thông tin được đề cập trong tài liệu. Việc thiếu trích dẫn chính xác có thể dẫn đến sự nhầm lẫn hoặc hiểu sai, ảnh hưởng đến độ tin cậy của nghiên cứu (Ali & Richardson, 2021). Do đó, một hệ thống giúp các nhà nghiên cứu tìm kiếm một tập hợp các tài liệu liên quan chặt chẽ đến đoạn văn họ đang làm việc sẽ rất hữu ích.

Nghiên cứu trình bày hệ thống khám phá trích dẫn sử dụng phương pháp truy xuất zero-shot. Hệ thống tận dụng mô hình ngôn ngữ lớn để trích xuất các bộ ba quan hệ ở cấp độ tài liệu từ một đoạn văn bản nhất định và truy xuất top-k tài liệu liên quan nhất. Từ các tài liệu được truy xuất này, thiết kế một tập hợp các câu lệnh để xác định trích dẫn chính xác cho đoạn văn bản đầu vào.

Để đánh giá hiệu quả của hệ thống, tiến hàng so sánh nó với các phương pháp cơ sở như TF-IDF và truy xuất vector dày đặc (Dense Vector Retrieval - DVR) trên tập dữ liệu SCIDOCA 2025 Shared Task 1. Kết quả cho thấy những cải thiện tích cực về điểm F1, chứng minh rằng việc tích hợp trích xuất quan hệ với truy xuất trích dẫn dựa trên mô hình ngôn ngữ lớn là một cách tiếp cận hiệu quả cho nhiệm vụ này.

II. Cơ sở lý thuyết

2.1. Truy xuất văn bản

Truy xuất văn bản là một vấn đề cơ bản trong lĩnh vực truy hồi thông tin, phát triển từ các phương pháp dựa trên từ khóa truyền thống đến các phương pháp học sâu phức tạp. Qua nhiều năm, các nhà nghiên cứu đã khám phá nhiều kỹ thuật khác nhau, bao gồm truy xuất dựa trên thuật ngữ, biểu diễn vector dày đặc và các hệ thống truy xuất dựa trên nơ-ron.

Truy xuất dựa trên thuật ngữ truyền thống: Một trong những cách tiếp cận sớm nhất và được sử dụng rộng rãi nhất đối với truy xuất văn bản là TF-IDF (Term Frequency-Inverse Document Frequency), xếp hạng tài liệu dựa trên tầm quan trọng của từ khóa (Jones, 1972). Mô hình không gian vector (VSM) (Turney & Pantel, 2010) biểu diễn các tài liệu dưới dạng các vector nhiều chiều, cho phép truy xuất dựa trên độ tương đồng sử dụng độ tương đồng cosine. Ngoài ra, BM25 (Robertson & Zaragoza, 2009) đã giới thiệu việc xếp hạng xác suất bằng cách xem xét tần suất thuật ngữ, độ dài tài liệu và tần suất tài liệu nghịch đảo, cải thiện đáng kể hiệu suất truy xuất. Các phương pháp này vẫn được sử dụng rộng rãi do tính hiệu quả và khả năng giải thích của chúng.

Truy xuất dựa trên vector dày đặc: Khi các nhiệm vụ truy xuất văn bản trở nên phức tạp hơn, các nhà nghiên cứu bắt đầu khám phá các biểu diễn vector dày đặc. Các nhúng từ (word embeddings), chẳng hạn như Word2Vec (Rong, 2014), GloVe (Pennington, 2014), và FastText (Joulin, 2017), đã cải thiện việc truy xuất văn bản bằng cách nắm bắt các mối quan hệ ngữ nghĩa giữa các từ. Tuy nhiên, các nhúng tĩnh này gặp khó khăn với các biến thể ngữ cảnh. Sự ra đời của các nhúng dựa trên các biến đổi, như BERT (Sun, 2019) và RoBERTa (Zhuang, 2021), đã cho phép truy xuất đoạn văn dày đặc (Dense Passage Retrieval - DPR) (Karpukhin, 2020), sử dụng các mô hình bi-encoder để mã hóa truy vấn và tài liệu thành các vector dày đặc cho việc truy xuất tương đồng. Các

phương pháp này đã cải thiện đáng kể độ chính xác truy xuất so với các cách tiếp cận dựa trên thuật ngữ truyền thống.

Truy xuất nơ-ron và tái xếp hạng: Các mô hình truy xuất dựa trên mạng nơ-ron theo cơ chế đầu cuối, chẳng hạn như các mô hình truy xuất dựa trên T5 (Raffel, 2020) và Gated Transformer Retrieval (GTR) (Ni, 2021), tinh chỉnh các mô hình transformer cho các nhiệm vụ truy xuất tài liệu. Ngoài ra, các mô hình truy xuất được hưởng lợi từ các cơ chế tái xếp hạng, trong đó mô hình truy xuất bước đầu tạo ra các tài liệu ứng viên, theo sau là một cross-encoder xếp hạng lại các tài liệu dựa trên sự hiểu biết ngữ cảnh sâu hơn (Thakur, 2021). Các kỹ thuật tái xếp hạng nâng cao đáng kể độ chính xác của truy xuất, đặc biệt là trong việc trả lời câu hỏi miền mở truy xuất văn bản pháp lý.

Thế hệ tăng cường truy xuất (RAG): Việc tích hợp các mô hình ngôn ngữ lớn vào các hệ thống truy xuất văn bản đã dẫn đến những tiến bộ đáng kể trong cách thông tin được truy cập và xử lý. Một sự phát triển đáng chú ý trong lĩnh vực này là thế hệ tăng cường truy xuất (Retrieval-Augmented Generation - RAG), một kỹ thuật kết hợp sức mạnh của các hệ thống truy xuất với khả năng tạo sinh của các mô hình ngôn ngữ lớn (Lewis, 2020). Cách tiếp cận này nâng cao khả năng của mô hình trong việc truy cập và kết hợp thông tin bên ngoài, từ đó cải thiện hiệu suất của chúng trên các nhiệm vụ đòi hỏi nhiều kiến thức như truy xuất văn bản.

2.2. Trích xuất quan hệ cấp tài liệu

Các mô hình trích xuất quan hệ truyền thống tập trung vào các quan hệ cấp câu, nhưng nhiều ứng dụng thực tế yêu cầu trích xuất quan hệ cấp tài liệu (DocRE). Trong bối cảnh này, các quan hệ được trích xuất qua nhiều câu, làm cho nhiệm vụ trở nên khó khăn hơn đáng kể do việc giải quyết đồng tham chiếu và liên kết thực thể (Zhang, 2023). Các mô hình dựa trên đồ thị và lập luận đa bước (multi-hop reasoning) đã được đề xuất để giải quyết những thách thức này bằng cách kết hợp ngữ cảnh tài liệu toàn cục (Nan, 2020).

2.3. Bài toán tìm kiếm trích dẫn

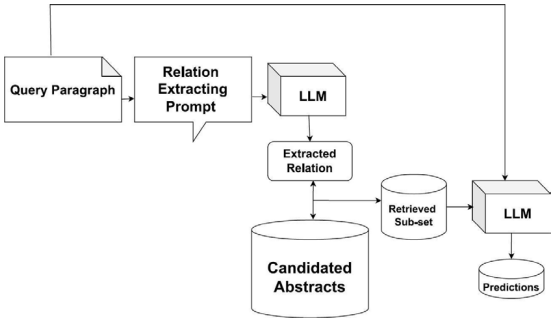
Mục tiêu của nhiệm vụ 1 thuộc tổ chức SCIDOCA là dự đoán chính xác tài liệu tham khảo cho một đoạn văn bản mà không cần xác định vị trí trích dẫn cụ thể trong đoạn. Với dữ liệu đầu vào gồm một đoạn văn truy vấn và một tập các tóm tắt ứng viên, hệ thống có nhiệm vụ định vị đúng mã định danh của bài báo đích.

Đặc thù của bài toán này đặt ra hai thách thức chính. Thứ nhất, độ dài lớn của cả đoạn văn truy vấn và các tóm tắt ứng viên làm gia tăng đáng kể lượng nhiễu thông tin trong quá trình truy xuất. Thứ hai, nghiên cứu của ứng viên chứa nhiều tài liệu gây nhiễu mạng độ tương đồng ngữ nghĩa cao với tài liệu đích, làm tăng độ phức tạp cho việc phân loại. Do đó, để đạt hiệu năng cao, phương pháp truy xuất phải đáp ứng đồng thời hai tiêu chuẩn cốt lõi: khả năng bao quát ngữ cảnh của văn bản dài và năng lực nhận diện các sắc thái ngữ nghĩa tinh vi nhằm phân biệt giữa các tài liệu có độ tương đồng cao. Hiệu năng của hệ thống được đánh giá bằng độ đo F1 nhằm đảm bảo sự cân bằng giữa độ chính xác và độ phủ của kết quả dự đoán.

III. Phương pháp nghiên cứu

Phương pháp truy xuất tăng cường dựa trên quan hệ nhằm giải quyết bài toán khám phá trích dẫn khoa học. Điểm cốt lõi của phương pháp đề xuất nằm ở việc khai thác các đặc trưng quan hệ được trích xuất

trực tiếp từ văn bản truy vấn, qua đó tối ưu hóa độ chính xác trong quá trình truy hồi thông tin. Kiến trúc tổng thể của hệ thống được minh họa chi tiết trong hình 1. Giai đoạn đầu tiên tiếp nhận một đoạn văn truy vấn đóng vai trò làm ngữ cảnh tìm kiếm. Nhận thấy đặc thù của văn bản khoa học thường bao hàm nhiều khái niệm và các mạch nghiên cứu đan xen, thực hiện thiết kế một câu lệnh chuyên biệt nhằm điều hướng mô hình ngôn ngữ lớn. Cụ thể, mô hình thực hiện nhiệm vụ trích xuất các bộ ba quan hệ có cấu trúc cùng các khái niệm cốt lõi, từ đó ánh xạ chính xác các liên kết bản chất giữa các thực thể trong văn bản. Việc mô hình hóa dữ liệu đầu vào dưới dạng cấu trúc hóa này cho phép cơ chế truy xuất được dẫn hướng bởi các ràng buộc ngữ nghĩa sâu, khắc phục triệt để hạn chế của các kỹ thuật so sánh từ khóa.



Hình 1: Tổng quan về hệ thống đề xuất

Tập biểu diễn quan hệ sau khi trích xuất được sử dụng làm cơ sở để truy hồi một không gian con chứa các tóm tắt tiềm năng nhất từ kho dữ liệu ứng viên. Đây được xem là bước thiết yếu, bởi kho dữ liệu ứng viên vốn chứa đựng tỷ lệ lớn các mẫu nhiễu khó những tài liệu có độ trùng lặp từ vựng cao nhưng sai lệch về ngữ cảnh trích dẫn đích. Thông qua cơ chế chắt lọc dựa trên sự đồng nhất về quan hệ ngữ nghĩa, hệ thống có khả năng triệt tiêu phần lớn lượng nhiễu, qua đó khuếch đại xác suất định vị chính xác tài liệu tham chiếu.

Ở giai đoạn suy luận cuối cùng, một mô hình LLM thứ hai được triển khai để xác định kết quả trích dẫn. Mô hình này thực hiện cơ chế đối chiếu chéo giữa tập tóm tắt đã truy xuất và đoạn văn truy vấn gốc, thông qua việc đánh giá đồng thời độ tương đồng văn bản lẫn tính mạch lạc của cấu trúc quan hệ. Đầu ra của mô hình là mã định danh duy nhất của bài báo hội tụ đầy đủ các tiêu chí ngữ nghĩa nhất. Việc áp dụng chiến lược phân cấp hai bước tích hợp cơ chế trích xuất quan hệ vào quy trình truy xuất và xếp hạng buộc hệ thống phải tập trung vào các cấu trúc ngữ cảnh cốt lõi. Bằng cách giảm thiểu sự phụ thuộc vào độ tương đồng văn bản bề mặt, phương pháp đề xuất nâng cao rõ rệt năng lực phân biệt đối với các tóm tắt mang tính đánh lừa cao. Hiệu năng tổng thể của kiến trúc được lượng hóa thông qua độ đo F, bảo đảm sự tối ưu hóa đồng thời giữa độ chính xác và độ phủ trong khả năng dự đoán.

IV. Kết quả và thảo luận

Giá hiệu suất của hệ thống, tiến hành thực nghiệm trên 1.000 truy vấn được chọn ngẫu nhiên từ tập dữ liệu. Điểm F1 được sử dụng làm thước đo đánh giá và so sánh phương pháp của mình với các phương pháp cơ sở sử dụng TF-IDF và Truy xuất vector dày đặc. Trong quá trình thực nghiệm, đã chọn mô hình “mistralai/Mistral-7B-Instruct-v0.3”cho cả việc trích xuất quan hệ và truy xuất trích dẫn chính xác. Đối với bước truy xuất đầu tiên, thu thập 20 tài liệu liên quan nhất từ kho tài liệu ứng viên.

Kết quả thực nghiệm được trình bày trong bảng 1 và bảng 2 thể hiện hiệu quả của các phương pháp truy xuất khác nhau và tác động của chúng đến hiệu suất dự đoán trích dẫn.

Từ bảng 1, chúng ta thấy rằng các phương pháp truy xuất truyền thống như TF-IDF và truy xuất dày đặc đạt điểm độ phủ cao, nhưng độ chính xác của chúng tương đối thấp. Điều này chỉ ra rằng mặc dù các phương pháp này truy xuất thành công tài liệu chính xác trong nhiều trường hợp, chúng cũng trả về một lượng lớn tài liệu không liên quan, do đó làm giảm độ chính xác. Như dự đoán, việc tăng số lượng tài liệu được truy xuất dẫn đến tăng độ phủ nhưng phải đánh đổi bằng việc độ chính xác giảm. Phương pháp truy xuất truyền thống hoạt động tốt nhất, dense retrieval-10, đạt điểm F1 là 0.3354, vượt trội hơn các phương pháp dựa trên TF-IDF.

Ngược lại, cách tiếp cận truy xuất dựa trên quan hệ đã thể hiện sự cải thiện đáng kể về độ chính xác (ví dụ: 0.8506 cho top-10). Tuy nhiên, điều này đi kèm với chi phí là độ phủ giảm, cho thấy rằng việc trích xuất quan hệ lọc ra nhiều tài liệu không liên quan nhưng cũng có thể loại bỏ một số tài liệu liên quan. Sự đánh đổi này thể hiện rõ khi so sánh phương pháp này với truy xuất dày đặc: trong khi truy xuất dựa trên quan hệ cung cấp một tập hợp tài liệu ứng viên chính xác hơn, nó không nhất thiết tối đa hóa độ phủ.

Bảng 1: So sánh hiệu suất của các phương pháp truy xuất khác nhau. Mỗi phương pháp được thực hiện với các tài liệu liên quan nhất từ các top-k khác nhau

Phương pháp	Recall	Precision	F1 Score
TF-IDF-10	0.8259	0.2060	0.3297
TF-IDF-15	0.9115	0.1715	0.2886
TF-IDF-20	0.9533	0.1534	0.2643
Dense retrieval-10	0.8401	0.2095	0.3354
Dense retrieval-15	0.9215	0.1733	0.2918
Dense retrieval-20	0.9615	0.1547	0.2665
Relation-based-10	0.8506	0.2156	0.3440
Relation-based-15	0.9300	0.1772	0.2976
Relation-based-20	0.9670	0.1576	0.2711

Bảng 2 thể hiện việc tích hợp LLM với truy xuất giúp tăng cường độ phủ đáng kể so với điểm truy xuất thô. Suy luận LLM với truy xuất dựa trên quan hệ đạt điểm F1 là 0.2912, cho thấy hiệu suất cân bằng tương tự như truy xuất dày đặc nhưng với sự đánh đổi độ phủ-độ chính xác khác nhau. Đáng chú ý, sự kết hợp giữa LLM với TF-IDF dẫn đến độ chính xác tốt hơn một chút so với LLM với truy xuất dày đặc, nhưng độ phủ của nó vẫn thấp hơn.

Bảng 2: So sánh hiệu suất của các phương pháp suy luận LLM

Phương pháp	Recall	Precision	F1 Score
LLM inference with TF-IDF	0.4079	0.2347	0.2980
LLM inference with Dense retrieval	0.3253	0.2590	0.2884
LLM inference with Relation-based	0.4626	0.2125	0.2912

Những phát hiện này nhấn mạnh tầm quan trọng của truy xuất dựa trên quan hệ trong việc tinh chỉnh các tài liệu ứng viên trước khi chuyển chúng cho LLM, đảm bảo rằng dự đoán cuối cùng không chỉ liên quan mà còn chính xác. Sự đánh đổi giữa độ chính xác và độ phủ là một cân nhắc chính trong việc thiết kế một hệ thống khám phá trích dẫn hiệu quả, và kết quả chỉ ra rằng sự kết hợp tối ưu giữa truy

xuất dày đặc cho độ phủ và lọc dựa trên quan hệ cho độ chính xác có thể mang lại hiệu suất tốt nhất.

V. Kết luận

Bài báo này đã trình bày một phương pháp truy xuất dựa trên quan hệ cho khám phá trích dẫn, tận dụng trích xuất quan hệ cấp tài liệu và các mô hình ngôn ngữ lớn để cải thiện độ chính xác dự đoán trích dẫn. Hệ thống trích xuất các đặc trưng quan hệ chính từ đoạn văn truy vấn để truy xuất một tập hợp con các tóm tắt ứng viên có ý nghĩa hơn, giảm nhiễu do các kỹ thuật truy xuất dựa trên văn bản truyền thống đưa vào.

Kết quả thực nghiệm chứng minh rằng trong khi truy xuất vector dày đặc đạt độ phủ cao, độ chính xác của nó vẫn tương đối thấp. Ngược lại, truy xuất dựa trên quan hệ cải thiện đáng kể độ chính xác, đảm bảo rằng chỉ các tóm tắt liên quan nhất mới được xem xét. Việc tích hợp suy luận LLM giúp tăng cường thêm độ phủ, dẫn đến một chiến lược truy xuất và xếp hạng cân bằng hơn. Sự kết hợp giữa truy xuất dựa trên quan hệ với suy luận LLM chứng tỏ là một cách tiếp cận hiệu quả cho các nhiệm vụ dự đoán trích dẫn.

Đối với các hướng nghiên cứu trong tương lai, hướng tới mục tiêu khám phá các mô hình truy xuất lại có thể điều chỉnh linh hoạt sự đánh đổi giữa độ phủ và độ chính xác dựa trên các đặc điểm của đoạn văn. Ngoài ra, việc kết hợp lập luận quan hệ dựa trên đồ thị có thể cải thiện hơn nữa hiệu quả truy xuất bằng cách nắm bắt các phụ thuộc ngữ cảnh sâu hơn giữa các tài liệu.

Tài liệu tham khảo

Ali, M. J., & Richardson, K. (2021). The impact of inaccurate citations in scientific writing: A systematic review. *Journal of Academic Writing, 11(2)*, 45-58.

Hassan, S., Mahdi, M. N., & Al-Jumeily, D. (2023). Citation recommendation systems: A comprehensive survey and future trends. *IEEE Access, 11*, 10234-10255.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11-21.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 427-431).

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459-9474.

Nan, G., Guo, Z., Sekulic, I., & Lu, W. (2020). Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1546-1557).

Ni, J., Qu, C., Lu, J., Dai, Z., Abrego, G. H., Ma, J., et al. (2021). Large dual encoders are generalizable retrievers. *arXiv*. https://arxiv.org/abs/2112.07899

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1-67.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval, 3*(4), 333-389.

Rong, X. (2014). *word2vec parameter learning explained*. arXiv. https://arxiv.org/abs/1411.2738

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics* (pp. 194-206). Springer.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot information retrieval evaluation. *arXiv*. https://arxiv.org/abs/2104.08663

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141-188.

Zhang, N., Yao, Y., Deng, S., Chen, X., Tan, C., Huang, M., et al. (2023). Document-level relation extraction: A survey. *ACM Transactions on Intelligent Systems and Technology*.

Zhuang, L., Wayne, L., Ya, S., & Jun, Z. (2021). A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics* (pp. 1218-1227).

RELATION-BASED ZERO-SHOT RETRIEVAL METHOD FOR CITATION DISCOVERY

Dao Xuan Phuc[1], Nguyen Thong Nhat[1], Nguyen Thanh Xuan[2]

Abstract: *The Citation Discovery Shared Task focuses on predicting the correct citation from a given candidate pool for a given input paragraph. The main challenges stem from the length of the abstracts and the high similarity among candidates, making it difficult to determine the exact paper to cite. To address this issue, this paper presents a system that retrieves the top-k most similar abstracts based on relational features extracted from the given text. From this subset, we leverage a Large Language Model (LLM) to accurately identify the most relevant citation. We evaluate our framework on the standard benchmark dataset, demonstrating the effectiveness of the proposed method in citation prediction.*

Keywords: *citation discovery, relation extraction, text retrieval, large language models, generative artificial intelligence*

---

[1] Faculty of Electric and Electronic Engineering, Hanoi Open University, Hanoi, Vietnam

[2] Peoples police university of technology and logistics, Hanoi, Vietnam